10

Culture and Research Infrastructure

Earlier chapters of this report have focused on what might be achieved experimentally and on the scientific and technical hurdles that must be overcome at the interface of biology and computing. This chapter focuses on the infrastructural underpinnings needed to support research at this interface. Note that because the influence of computing on biology has been much more significant than the influence of biology on computing, the discussion in this chapter is focused mostly on the former.

10.1 SETTING THE CONTEXT

In 1991, Walter Gilbert sketched a vision of 21st century biology (described in Chapter 1) and noted the changes in intellectual orientation and culture that would be needed to realize that vision. He wrote:

> To use [the coming] flood of [biological] knowledge [i.e., sequence information], which will pour across the computer networks of the world, biologists not only must become computer-literate, but also change their approach to the problem of understanding life. . . . The next tenfold increase in the amount of information in the databases will divide the world into haves and have-nots, unless each of us connects to that information and learns how to sift through it for the parts we need. This is not more difficult than knowing how to access the scientific literature as it is at present, for even that skill involves more than a traditional reading of the printed page, but today involves a search by computer. . . . We must hook our individual computers into the worldwide network that gives us access to daily changes in the database and also makes immediate our communications with each other. The programs that display and analyze the material for us must be improved—and we [italics added] must learn how to use them more effectively.[1]

In short, Gilbert pointed out the need for institutional change (in the sense of individual life scientists learning to cooperate with each other) and for biologists to learn how to use the new tools of information technology.

Because the BioComp interface encompasses a variety of intellectual paradigms and disparate institutions, Section 10.2 describes the organizational and institutional infrastructure supporting work at this interface, illustrating a variety of programs and training approaches. Section 10.3 addresses some

[1]W. Gilbert, "Toward a Paradigm Shift in Biology," *Nature* 349(6305):99, 1991.

of the barriers that affect research at the BioComp interface. Chapter 11 is devoted to proposing possible ways of helping to reduce the negative impact of these barriers.

10.2 ORGANIZATIONS AND INSTITUTIONS

Efforts to pursue research at the BioComp interface, as well as the parallel goal of attracting and training a sufficient workforce, are supported by a number of institutions and organizations in the public and private sectors. A prime mover is the U.S. government, both by pursuing research in its own laboratories, and by providing funding to other, largely academic, organizations. However, the government is only a part of a larger web of collaborating (and competing) academic departments, private research institutions, corporations, and charitable foundations.

10.2.1 The Nature of the Community

The members of an established scientific community can usually be identified by a variety of commonalities—fields in which their degrees were received, journals in which they publish, and so on.[2] The fact that important work at the BioComp interface has been undertaken by individuals who do not necessarily share such commonalities indicates that the field in question has not jelled into a single community, but in fact is composed of many subcommunities. Members of this community may come from any of a number of specialized fields, including (but not restricted to) biology, computer science, engineering, chemistry, mathematics, and physics. (Indeed, as the various epistemological and ontological discussions of previous chapters suggest, even philosophers and historians of science may have a useful role to play.)

Because the intellectual contours of work at the intersection have not been well established, the definition of the community must be broad and is necessarily somewhat vague. Any definition must encompass a multitude of cultures and types, leaving room for approaches that are not yet known. Furthermore, the field is sufficiently new that people may enter it at many different stages of their careers.

For perspective, it is useful to consider some possible historical parallels with the establishment of biochemistry, biophysics, and bioengineering as autonomous disciplines. In each case, the phenomena associated with life have been sufficiently complex and interesting to warrant the bringing to bear of specialized expertise and intellectual styles originating in chemistry, physics, and engineering. Nonbiologists, including chemists, physicists, and engineers, have made progress on some biologically significant problems precisely because their approaches to problems differed from those of biologists and thus have advanced biological understanding because they were not limited by what biologists felt could not be understood. On the other hand, chemists, physicists, and engineers have also pursued many false or unproductive lines of inquiry because they have not appreciated the complexity that characterizes many biological phenomena or because they addressed problems that biologists already regarded as solved. Eventually, biochemistry, biophysics, and bioengineering became established in their own right as education and cultural inculcation from both parent disciplines came to be required.

It is also to be expected that the increasing integration of computing and information into biology will raise difficult questions about the nature of biological research and science. If an algorithm to examine the phylogenetic tree of life is too slow to run on existing hardware, clearly a new algorithm must be developed. Does developing such an algorithm constitute biological research? Indeed, modern biology is sufficiently complex that many of the most important biological problems are not easily tamed by existing mathematical theory, computational models, or computing technologies. Ultimately, success in understanding biological phenomena will depend on the development and application of new tools throughout the research process.

---

[2]T.S. Kuhn, *The Structure of Scientific Revolutions,* Third Edition, University of Chicago Press, Chicago, IL, 1996.

10.2.2 Education and Training

Education, either formal or informal, is essential for practitioners of one discipline to learn about another, and there are many different venues in which training for the BioComp interface may occur. (Contrast this to a standard program in physics, for example, in which a very typical career path involves an undergraduate major in physics, graduate education in physics culminating in a doctorate, and a postdoctoral appointment in physics.)

Reflecting this diversity, it is difficult to generalize about approaches toward academic training at the BioComp interface, since different departments and institutions approach it with varied strategies. One main difference in approaches is whether the initiative for creating an educational program and the oversight and administration of the program come from the computer science department or the biology department. Other differences include whether it is a stand-alone program or department, or a concentration or interdisciplinary program that requires a student or researcher to have a "home" department as well, and whether the program was established primarily as a research program for postdoctoral fellows and professors (and is slowly trickling down to undergraduate and graduate education), or as an undergraduate curriculum that is slowly building its way up to a research program. Those differences in origin result in varying emphases on what constitutes core subject matter, whether interdisciplinary work is encouraged and how it is handled, and how research is supported and evaluated.

What is clear is that this is an active area of development and investment, and many major colleges and universities have a formal educational program of some sort at the BioComp interface (generally in bioinformatics or computational biology) or are in the process of developing one. Of course, there is not yet widespread agreement on what the curriculum for this new course of study should be[3] or indeed if there should be a single, standard, curriculum.

10.2.2.1 General Considerations

As a general rule, serious work at the BioComp interface requires knowledge of both biology and computing. For example, many models and simulations of biological phenomena are constrained by lack of quantitative data. The paucity of measurements of in vivo rates or parameters associated with dynamics means that it is difficult to understand systems from a dynamic, rather than a static, point of view. For example, to further the use of biological modeling and simulation, kinetics should be an important part of early biological courses, including biochemistry and molecular biology, to instill an appreciation in experimental biologists that kinetics is important. The requisite background in quantitative methods is likely to include some nontrivial exposure to continuous mathematics, nonlinear dynamics, linear algebra, probability and statistics, as well as computer programming and algorithm design.

From the engineering side, few nonbiologists get any exposure to biological laboratory research or develop an understanding of the collection and analysis of biological data. This also leads to unrealistic expectations of what can be done practically, how repeatable (or unrepeatable) a set of experiments can be, and how difficult it can be to understand the system in detail. Computer scientists also require exposure to probability, statistics, laboratory technique, and experimental design in order to understand the biologist's empirical methodology. More fundamentally, nonbiologists working at the BioComp interface must have an understanding of the basic principles relevant to the biological problem domains of interest, such as physiology, phylogeny, or proteomics. (A broad perspective on biology, including some exposure to evolution, ecosystems, and metabolism, is certainly desirable, but is likely not absolutely necessary.)

Finally, it must be noted that many students choose to study biology because it is a science whose study has traditionally not involved mathematics to any significant extent. Similarly, W. Daniel Hillis

[3]R. Altman, "A Curriculum for Bioinformatics: The Time Is Ripe," *Bioinformatics* 14(7):549-550, 1998.

has noted that "biologists are biologists because they love living things. A computation is not alive."[4] Indeed, this has been true for several generations of students, so that many of these same students are now incumbent instructors of biology. Managing this particular problem will pose many challenges.

10.2.2.2 Undergraduate Programs

The primary rationale for undergraduate programs at the BioComp interface is that the undergraduate years of university education in the sciences carry the greatest burden in teaching a student the professional language of a science and the intellectual paradigms underlying the practice of that science. The term "paradigm" is used here in the original sense first expounded by Kuhn, which includes the following:[5]

- Symbolic generalizations, which the community uses without question,
- Beliefs in particular models, which help to determine what will be accepted as an explanation or a solution,
- Values concerning prediction (e.g., predictions must be accurate, quantitative) and theories (e.g., theories must be simple, self-consistent, plausible, compatible with other theories in current use), and
- Exemplars, which are the concrete problem solutions that students encounter from the start of their scientific education.

The description in Section 10.3.1 suggests that the disciplinary paradigm of biology is significantly different from that of computer science. Because the de novo learning of one paradigm is easier than subsequently learning a second paradigm that may (apparently) be contradictory or incommensurate with one that has already been internalized, the argument for undergraduate exposure is based on the premise that simultaneous exposure to the paradigms of two disciplines will be more effective than sequential exposure (as would be the case for someone receiving an undergraduate degree in one field and then pursuing graduate work in another).

Undergraduate programs in most scientific courses of study are generally designed to prepare students for future academic work in the field. Thus, the goal of undergraduate curricula at the BioComp interface is to expose students to a wide range of biological knowledge and issues and to the intellectual tools and constructs of computing such as programming, statistics, algorithm design, and databases. Today, most such programs focus on bioinformatics or computational biology, and in the most typical cases, the integration of biology and computing occurs later rather than earlier in these programs (e.g., as senior-year capstone courses).

Individual programs vary enormously in the number of computer science classes required. For example, the George Washington University Department of Computer Science offers a concentration in bioinformatics leading to a B.S. degree; the curriculum includes 17 computer science courses and 4 biology courses, plus a single course on bioinformatics. The University of California, Los Angeles (UCLA) program in cybernetics offers a concentration in bioinformatics, in contrast, in which the students can take as few as seven computer science courses, including four programming classes and two biology-themed classes. In other cases, a university may have an explicit undergraduate major in bioinformatics associated with a bioinformatics department. Such programs are traditionally structured in the sense of having a set of specific courses required for matriculation.

In addition to concentrations at the interface, a number of other approaches have been used to prepare undergraduates:

- An explicitly interdisciplinary B.S. science program can expose students to the interrelationships of the basic sciences. Sometimes these are co-taught as single units: students in their first year may take

[4]W.D. Hillis, "Why Physicists Like Models, and Biologists Should," *Current Biology* 3(2):79-81, 1993.
[5]T.S. Kuhn, *The Structure of Scientific Revolutions*, Third Edition, University of Chicago Press, Chicago, 1996.

mathematics, physics, chemistry, and biology as a block, taught by a team of dedicated professors. Modules in which ideas from one discipline are used to solve problems in another are developed and used as case studies for motivating the connections between the topics. Other coordinated science programs intersperse traditional courses in the disciplines with co-taught interdisciplinary courses (Examples: Applications of Physical Ideas in Biological Systems; Dimensional Analysis in the Sciences; Mathematical Biology).

- A broad and unrestricted science program can allow students to count basic courses in any department toward their degree or to design and propose their personal degree program. Such a system gives graduates an edge in the ability to transcend boundaries between disciplines. A system of co-advising to help students balance needs with interests would be vital to ensure that such open programs function well.
- Courses in quantitative science with explicit ties to biology may be more motivating to biology students. Some anecdotal evidence indicates that biology students can do better in math and physics when the examples are drawn from biology; at the University of Washington, the average biology student's grade in calculus rose from C to B+ when "calculus for biologists" was introduced.[6] (Note that such an approach requires that the instructor have the knowledge to use plausible biological examples—a point suggesting that simply handing off modules of instruction will not be successful.)
- Summer programs for undergraduates offer undergraduates an opportunity to get involved in actual research projects while being exposed to workshops and tutorials in a range of issues at the BioComp interface. Many such programs are funded by a National Science Foundation or National Institutes of Health program.[7]

When none of these options are available, a student can still create a program informally (either on his or her initiative or with the advice and support of a sympathetic faculty member). Such a program would necessarily include courses sufficient to impart a thorough quantitative background (mathematics, physics, computer science) as well as a solid understanding of biology. As a rule, quantitative training should come first, because it is often difficult to develop expertise in quantitative approaches later in the undergraduate years. Exposure to intriguing ideas in biology (e.g., in popular lecture series) would also help to encourage interest in these directions.

Finally, an important issue at some universities is the fact that computer science departments and biology departments are located in different schools (school of engineering versus school of arts and sciences). As a result, biology majors may well face impediments to enrolling in courses intended for computer science majors, and vice versa. Such a structural impediment underlines both the need and the challenges for establishing a biological computing curriculum.

10.2.2.3 The BIO2010 Report

In July 2003, the National Research Council (NRC) released *Bio 2010: Undergraduate Education to Prepare Biomedical Research Scientists* (National Academies Press, Washington, DC). This report concluded that undergraduate biology education had not kept pace with computationally driven changes in life sciences research, among other changes, and recommended that mathematics, physics, chemistry, computer science, and engineering be incorporated into the biology curriculum to the point that interdisciplinary thinking and work become second nature for biology students. In particular, the report noted "the importance of building a strong foundation in mathematics, physical and information sciences to prepare students for research that is increasingly interdisciplinary in character."

The report elaborated on this point in three other recommendations—that undergraduate life sci-

[6]Mary Lidstrom, University of Washington, personal communication, August 1, 2003.
[7]See http://www.nsf.gov/pubs/2002/nsf02109/nsf02109.htm.

ence majors should be exposed to engineering principles and analysis, should receive quantitative training in a manner integrated with biological content, and should develop enough familiarity with computer science that they can use information technology effectively in all aspects of their research.

10.2.2.3.1 Engineering In arguing for exposure to engineering, the report noted that the notion of function (of a device or organism) is common to both engineering and biology, but not to mathematics, physics, or chemistry. Echoing the ideas described in Chapter 6 of this report, BIO2010 concluded:

> Understanding function at the systems level requires a way of thinking that is common to many engineers. An engineer takes building blocks to build a system with desired features (bottom-up). Creating (or re-creating) function by building a complex system, and getting it to work, is the ultimate proof that all essential building blocks and how they work in synchrony are truly understood. Getting a system to work typically requires (a) an understanding of the fundamental building blocks, (b) knowledge of the relation between the building blocks, (c) the system's design, or how its components fit together in a productive way, (d) system modeling, (e) construction of the system, and (f) testing the system and its function(s). . . . Organisms can be analyzed in terms of subsystems having particular functions. To understand system function in biology in a predictive and quantitative fashion, it is necessary to describe and model how the system function results from the properties of its constituent elements.

The pedagogical conclusion was clear in the report:

> Understanding cells, organs, and finally animals and plants at the systems level will require that the biologist borrow approaches from engineering, and that engineering principles are introduced early in the education of biologists. . . . Students should be frequently confronted throughout their biology curriculum with questions and tasks such as how they would design 'xxx,' and how they would test to see whether their conceptual design actually works. [For example,] they should be asked to simulate their system, determine its rate constants, determine regimes of stability and instability, investigate regulatory feedback mechanisms, and other challenges.

A second dimension in which engineering skills can be useful is in logistical planning. There are many areas in biology now where it is relatively easy to conceive of an important experiment, but drawing out the implications of the experiment involves a combinatorial explosion of analytical effort and thus is not practical to carry out. It is entirely plausible that many important biological discoveries will depend on both the ability to conceive an experiment and the ability to reconceive and restructure it logistically so that it is, in fact, doable. Engineers learn to apply their fundamental scientific knowledge in an environment constrained by nonscientific concerns, such as cost or logistics, and this ability will be critically important for the biologist who must undertake the restructuring described above. Box 10.1 provides a number of examples of engineering for life science majors.

10.2.2.3.2 Quantitative Training In its call for greater quantitative training, the BIO2010 report echoed that of other commentators.[8] Recognizing that quantitative analysis, modeling, and prediction play important roles in today's biomedical research (and will do so increasingly in the future), the report noted the importance to biology students of understanding concepts such as rate of change, modeling, equilibrium, and stability, structure of a system, and interactions among components, and argued that every student should acquire the ability to analyze issues arising in these contexts in some depth, using analytical methods (including paper-and-pencil techniques) and appropriate computational tools. As part of a necessary background, the report suggested that an appropriate course of study would include aspects of probability, statistics, discrete models, linear algebra, calculus and differential equations, modeling, and programming (Box 10.2).

[8]See for example, A. Hastings and M.A. Palmer, "A Bright Future for Biologists and Mathematicians," *Science* 299(5615):2003-2004, 2003, available at http://www.biosino.org/bioinformatics/a%20bright%20future.pdf.

Box 10.1
Engineering for Life Science Majors

One example of an engineering topic suitable for inclusion in a biology curriculum is the subject of long-range neuron signals. Introducing such a topic might begin with the electrical conductivity of salt water and of the lipid cell membrane, and the electrical capacitance of the cell membrane. It would next develop the simple equations for the attenuation of a voltage applied across the membrane at one end of an axon "cylinder" with distance down the axon, and the effect of membrane capacitance on signal dynamics for time-varying signals.

After substituting numbers, it becomes clear that amplifiers will be essential. On the other hand, real systems are always noisy and imperfect; amplifiers have limited dynamical range; and the combination of these facts makes sending an analog voltage signal through a large number of amplifiers essentially impossible.

The pulse coding of information overcomes the limitations of analog communication. How are "pulses" generated by a cell? This would lead to the power supply needed by an amplifier—ion pumps and the Nernst potential. How are action potentials generated? A first example of the transduction of an analog quantity into pulses might be stick-slip fraction, in which a block resting on a table, and pulled by a weak spring whose end is steadily moved, moves in "jumps" whose distance is always the same. This introduction to nonlinear dynamics contains the essence of how an action potential is generated.

The "negative resistance" of the sodium channels in a neuron membrane provides the same kind of "breakdown" phenomenon. Stability and instabilities (static and dynamic) of nonlinear dynamical systems can be analyzed, and finally the Hodgkin-Huxley equations illustrated.

The material is an excellent source of imaginative laboratories involving electrical measurements, circuits, dynamical systems, batteries and the Nernst potential, information and noise, and classical mechanics. It has great potential for simulations of systems a little too complicated for complete mathematical analysis, and thus is ideal for teaching simulation as a tool for understanding.

Other biological phenomena that can be analyzed using an engineering approach and that are suitable for inclusion in a biology curriculum include the following:

- The blood circulatory system and its control; fluid dynamics; pressure and force balance;
- Swimming, flying, walking, dynamical description, energy requirements, actuators, control; material properties of biological systems and how their structure relates to their function (e.g., wood, hair, cell membrane cartilage);
- Shapes of cells: force balance, hydrostatic pressure, elasticity of membrane and effects of the spatial dependence of elasticity; effects of cytoskeletal force on shape; and
- Chemical networks for cell signaling; these involve the concepts of negative feedback, gain, signal-to-noise, bandwidth, and cross-talk. These concepts are simple to experience in the context of how an electrical amplifier can be built from components.

SOURCE: Adapted from National Research Council, *BIO2010: Transforming Undergraduate Education for Future Research Biologists*, The National Academies Press, Washington, DC, 2003.

10.2.2.3.3 Computer Science Finally, the BIO2010 report noted the importance of information technology-based tools for biologists. It recommended that all biology majors be able to develop simulations of physiological, ecological, and evolutionary processes; to modify existing applications as appropriate; to use computers to acquire and process data; to carry out statistical characterization of the data and perform statistical tests; to graphically display data in a variety of representation; and to use information technology (IT) to carry out literature searches, locate published articles, and access major data-

Box 10.2
Essential Concepts of Mathematics and Computer Science for Life Scientists

Calculus

- Complex numbers
- Functions
- Limits
- Continuity
- The integral
- The derivative and linearization
- Elementary functions
- Fourier series
- Multidimensional calculus: linear approximations, integration over multiple variables

Linear Algebra

- Scalars, vectors, matrices
- Linear transformations
- Eigenvalues and eigenvectors
- Invariant subspaces

Dynamical Systems

- Continuous time dynamics—equations of motion and their trajectories
- Test points, limit cycles, and stability around them
- Phase plane analysis
- Cooperativity, positive feedback, and negative feedback
- Multistability
- Discrete time dynamics—mappings, stable points, and stable cycles
- Sensitivity to initial conditions and chaos

Probability and Statistics

- Probability distributions
- Random numbers and stochastic processes
- Covariation, correlation, and independence
- Error likelihood

Information and Computation

- Algorithms (with examples)
- Computability
- Optimization in mathematics and computation
- "Bits": information and mutual information

Data Structures

- Metrics: generalized "distance" and sequence comparisons
- Clustering
- Tree relationships
- Graphics: visualizing and displaying data and models for conceptual understanding

SOURCE: Reprinted from National Research Council, *BIO2010: Transforming Undergraduate Education for Future Research Biologists*, The National Academies Press, Washington, DC, 2003.

bases. From the perspective of this report, Box 10.3 describes some of the essential intellectual aspects of computer science that biologists must understand.

Recognizing that students might require competence at multiple levels depending on their needs, the BIO2010 report identified three levels of competence as described in Box 10.4.

Box 10.3
Essential Concepts of Computer Science for the Biologist

Key for the computer scientist is the notion of a field that focuses on information, on understanding of computing activities through mathematical and engineering models and based on theory and abstraction, on the ways of representing and processing information, and on the application of scientific principles and methodologies to the development and maintenance of computer systems—whether they are composed of hardware, software, or both.

There are many views of understanding the essential concepts of computer science. One view, developed in 1991 in the NRC report *Computing the Future*, is that the key intellectual themes in computing are algorithmic thinking, the representation of information, and computer programs.[1]

- An algorithm is an unambiguous sequence of steps for processing information. Of particular relevance is how the speed of the algorithm varies as a function of problem size—the topic of algorithmic complexity. Typically, a result from algorithmic complexity will indicate the scaling relationships between how long it takes to solve a problem and the size of the problem when the solution of the problem is based on a specific algorithm. Thus, algorithm A might solve a problem in a time of order N^2, which means that a problem that is 100 times as large would take $100^2 = 10{,}000$ times as long to solve, whereas a faster algorithm B might solve the same problem in time of order $N \ln N$, which means a problem 100 times as large would take $100 \ln 100 = 460.5$ times as long to solve. Such results are important because all computer programs embed algorithms within them. Depending on the functional relationship between run time and problem size, a given program that works well on a small set of test data may—or may not—work well (run in a reasonable time) for a larger set of real data. Theoretical computer science thus imposes constraints on real programs that software developers ignore at their own peril.
- The representation of information or a problem in an appropriate manner is often the first step in designing an algorithm, and the choice of one representation or another can make a problem easy or difficult, and its solution slow or fast. Two issues arise: (1) how should the abstraction be represented, and (2) how should the representation be structured properly to allow efficient access for common operations? For example, a circle of radius 2 can be represented by an equation of the form $x^2 + y^2 = 4$ or as a set of points on the circle ((0.00, 2.00), (0.25, 1.98), (0.50, 1.94), (0.75, 1.85), (1.00, 1.73), (1.25, 1.56), (1.50, 1.32), (1.75, 0.97), (2.00, 0.00)), and so on. Depending on the purpose, one or the other of these representations may be more useful. If the circle of radius 2 is just a special case of a problem in which circles of many different radii are involved, representation as an equation may be more appropriate. If many circles of radius 2 have to be drawn on a screen and speed is important, a listing of the points on the circle may provide a faster basis for drawing such circles.
- A computer program expresses algorithms and structure information using a "programming language." Such languages provide a way to represent an algorithm precisely enough that a "high-level" description (i.e., one that is easily understood by humans) can be translated mechanically ("compiled") into a "low-level" version that the computer can carry out ("execute"); the execution of a program by a computer is what allows the algorithm to be realized tangibly, instructing the computer to perform the tasks the person has requested. Computer programs are thus the essential link between intellectual constructs such as algorithms and information representations and the computers that perform useful tasks.

[1]The discussion below is adapted from Computer Science and Telecommunications Board, National Research Council, *Computing the Future: A Broader Agenda for Computer Science and Engineering*, National Academy Press, Washington, DC, 1992.

continued

Box 10.3 Continued

This last point is often misunderstood. For many outsiders, computer science is the same as computer programming—a view reinforced by many introductory "computer science" courses that emphasize the writing of computer programs. But it is better to understand computer programs as the specialized medium in which the ideas and abstractions of computer science are tangibly manifested. Focusing on the writing of the computer program without giving careful consideration to the abstractions embodied in the program is not unlike understanding the writing of a novel as no more than the rules of grammar and spelling.

Algorithmic thinking, information representation, and computer programs are themes central to all subfields of computer science and engineering research. They also provide material for intellectual study in and of themselves, often with important practical results. The study of algorithms is as challenging as any area of mathematics, and one of practical importance as well, since improperly chosen or designed algorithms may solve problems in a highly inefficient manner. The study of programs is a broad area, ranging from the highly formal study of mathematically proving programs correct to very practical considerations regarding tools with which to specify, write, debug, maintain, and modify very large software systems (otherwise called software engineering). Information representation is the central theme underlying the study of data structures (how information can best be represented for computer processing) and much of human-computer interaction (how information can best be represented to maximize its utility for human beings).

Finally, computer science is closely tied to an underlying technological substrate that evolves rapidly. This substrate is the "stuff" out of which computational hardware is made, and the exponential growth that characterizes its evolution makes it possible to construct ever-larger, ever-more-complex systems—systems that are not predictable based on an understanding of their individual components. (As one example, the properties of the Internet prove a rich and surprisingly complex area of study even though its components—computers, routers, fiber-optic cables—are themselves well understood.)

A second report of the National Research Council described fluency with information technology as requiring three kinds of knowledge: skills in using contemporary IT, foundational concepts about IT and computing, and intellectual capabilities needed to think about and use IT for purposeful work.[2] The listing below is the perspective of this report on essential concepts of IT for everyone:

- *Computers* (e.g., programs as a sequence of steps, memory as a repository for program and data, overall organization, including relationship to peripheral devices).
- *Information systems* (e.g., hardware and software components, people and processes, interfaces (both technology interfaces and human-computer interfaces), databases, transactions, consistency, availability, persistent storage, archiving, audit trails, security and privacy and their technological underpinnings).
- *Networks:* physical structure (messages, packets, switching, routing, addressing, congestion, local area networks, wide area networks, bandwidth, latency, point-to-point communication, multicast, broadcast, Ethernet, mobility), and logical structure (client/server, interfaces, layered protocols, standards, network services).
- *Digital representation of information:* concept of information encoding in binary form; different information encodings such as ASCII, digital sound, images, and video/movies; precision, conversion and interoperability (e.g., of file formats), resolution, fidelity, transformation, compression, and encryption; standardization of representations to support communication.
- *Information organization* (including forms, structure, classification and indexing, searching and retrieving, assessing information quality, authoring and presentation, and citation; search engines for text, images, video, audio).
- *Modeling and abstraction:* methods and techniques for representing real-world phenomena as computer models, first in appropriate forms such as systems of equations, graphs, and relationships, and then in appropriate programming objects such as arrays or lists or procedures. Topics include continuous and discrete

[2]Computer Science and Telecommunications Board, National Research Council, *Being Fluent with Information Technology*, National Academy Press, Washington, DC, 1999.

models, discrete time events, randomization, and convergence, as well as the use of abstraction to hide irrelevant detail.
- *Algorithmic thinking and programming:* concepts of algorithmic thinking, including functional decomposition, repetition (iteration and/or recursion), basic data organization (record, array, list), generalization and parameterization, algorithm vs. program, top-down design, and refinement.
- *Universality and computability:* ability of any computer to perform any computational task.
- *Limitations of information technology:* notions of complexity, growth rates, scale, tractability, decidability, and state explosion combine to express some of the limitations of information technology; connections to applications, such as text search, sorting, scheduling, and debugging.
- *Societal impact of information and information technology:* technical basis for social concerns about privacy, intellectual property, ownership, security, weak/strong encryption, inferences about personal characteristics based on electronic behavior such as monitoring Web sites visited, "netiquette," "spamming," and free speech in the Internet environment.

A third perspective is provided by Steven Salzberg, senior director of bioinformatics at the Institute for Genomic Research in Rockville, Maryland. In a tutorial paper for biologists, he lists the following areas as important for biologists to understand:[3]

- Basic computational concepts (algorithms, program execution speed, computing time and space requirements as a function of input size; really expensive computations),
- Machine learning concepts (learning from data, memory-based reasoning),
- Where to store learned knowledge (decision trees, neural networks),
- Search (defining a search space, search space size, tree-based search),
- Dynamic programming, and
- Basic statistics and Markov chains.

[3]S.L. Salzberg, "A Tutorial Introduction to Computation for Biologists," *Computational Methods in Molecular Biology,* S.L. Salzberg, D. Searls, and S. Kasif, eds., Elsevier Science Ltd., New York, 1998.

10.2.2.4 Graduate Programs

Graduate programs at the BioComp interface are often intended to provide B.S. graduates in one discipline with the complementary expertise of the other. For example, individuals with bachelor's degrees in biology may acquire computational or analytical skills during early graduate school, with condensed "retraining" programs that expose then to nonlinear dynamics, algorithms, and so on. Alternatively, individuals with bachelor's degrees in computer science might take a number of courses to expose them to essential biological concepts and techniques.

Graduate education at the interface is much more diverse than at the undergraduate level. Although there is general agreement that an undergraduate degree should expose the student to the component sciences and prepare him or her for future work, the graduate degree involves a far wider array of goals, focuses, fields, and approaches. Like undergraduate programs, graduate programs can be stand-alone departments, independent interdisciplinary programs, or certificate programs that require students to have a "home" department.

A bioinformatics program oriented toward genomics is very common. Virginia Tech's program, for example, has been renamed the program in "Genetics, Bioinformatics, and Computational Biology," indicating its strong focus on genetic analysis. In contrast, the Keck Graduate Institute at Claremont stresses the interdisciplinary skill set necessary for the effective management of companies that straddle the biology-quantitative science boundary. It awards a master's of bioscience, a professional degree

Box 10.4
Competence and Expertise in Computer Science for Biology Students

The BIO2010 report recommended that all biology students receive instruction in computer science, distinguishing among three levels of competency. From lowest to highest, these include the following:

- *Fluency*. Based on the NRC report *Being Fluent with Information Technology,* fluency refers to the ability of biology students to use information technology today and to adapt to changes in IT in the future. For example, they need a basic understanding of how computers work and of programming, and a higher degree of fluency in using networks and databases. Students should also be exposed to laboratory experiences using MEDLINE, GenBank, and other biological databases, as well as physiological and ecological simulations. For example, students could be asked to use computer searches to track down all known information about a given gene and the protein it encodes, including both structure and function. This would involve exploring the internal structure of the gene (exons, introns, promoter, transcription factor binding sites); the regulatory control of the gene; sequence homologues of the gene and the protein; the structure and function of the protein; gene interaction networks and metabolic pathways involving the protein; and interactions of the protein with other proteins and with small molecules.
- *Capability in program design for computational biology and genomics applications*. Students at this level acquire the minimal skills required to be effective computer users within a computationally oriented biology research team. For example, they would learn structured software development and selected principles of computer science, with applications in computational biology and allied disciplines, and would use examples and tutorials drawn from problems in computational biology.
- *Capability in developing software tools for use by the biology community*. At this sophisticated level, students need a grounding in discrete mathematics, data structures, and algorithms, as well as database management systems, information systems, software engineering, computer graphics, or computer simulation techniques. Students at this level would be able to design and specify database and information systems for use by the entire community. Of special interest will be tools that require background in graph theory, combinatorics, and computational geometry as applications in high-throughput genomics research and rational drug design become increasingly important.

SOURCE: Adapted from National Research Council, *BIO2010: Transforming Undergraduate Education for Future Research Biologists,* The National Academies Press, Washington, DC, 2003.

somewhat like an M.B.A. with a science requirement. Some programs, such as Stanford's, are administered by the medical school, leading to a focus on medical informatics as well as bioinformatics. This would include topics such as clinical trials and image analysis, which would not show up in a more traditional genomics-focused bioinformatics degree.

The Research Training Program of the Keck Center for Computational and Structural Biology is intended to develop one of two different kinds of expertise. Emerging from this program, a trainee would be a computational expert well versed in computer science and quantitative methods who would also be knowledgeable in at least one application area of biological significance, or an expert in some biological area (e.g., molecular biology) who would also be aware of the most advanced concepts in computing. Students entering from computational backgrounds take at least three courses in biology-biochemistry-biophysics areas, while students entering from biological backgrounds at least three courses in computational areas. In addition, all students take an introductory course in computational science. Dissertation research is supervised by a committee with faculty members as required by the student's home department, but with representation from the computational biology faculty at other Keck Center institutions as well. Research can be undertaken in areas including the visualization of biological complexes, the development of DNA and protein sequence analysis, and advanced simulations.

A challenge for a field as interdisciplinary as this is that incoming students will arrive with possibly completely non-overlapping backgrounds. Most programs accept a B.S. in computer science, biology, or math as a prerequisite; to produce a well-rounded computational biologist will require very different training programs. The University of Colorado's certificate program in computational biology requires incoming students to take preparatory classes in "Biology for Computer Scientists," "Computer Science for Bioscientists," or "Mathematics for Bioscientists," depending on what the student missed earlier in his or her education.

An advantage of graduate programs is that when communication among faculty of different disciplines is good, graduate projects provide an ideal opportunity for students to work in an interdisciplinary environment. In some cases, work with adjunct professors from industry can lead to exciting projects. On the other hand, if communication between faculty is poor (which may be possible for reasons described later in this chapter), a graduate student dependent on completing a project (e.g., a dissertation) can get caught in the middle of a dispute with no way to graduate.

10.2.2.5 Postdoctoral Programs

Postdoctoral programs at the BioComp interface are also varied. Some postgraduate programs are explicitly aimed at "conversion," that is, training a fully trained member of one field (usually biology) in the basic tenets of its complement. For example, the University of Pennsylvania's postdoctoral program in computational biology is a master's degree in computer and information systems, designed for those with Ph.D.s in biology who need the training. Other programs focus on involving the participant in research and laboratory work, in preparation for industry or a faculty position, just as in postdoctoral programs in other fields.

Some programs, such as Duke's Center for Bioinformatics and Computational Biology, are similar to graduate programs in that they focus on genome analysis. Others, like the Johns Hopkins' program in computational biology, are firmly grounded in genomics but are pointedly reaching out to larger questions of integrative biology and experimental biology.

In promoting postdoctoral programs at the interface of computing and biology, it will be necessary to take into account the very different traditions of the two fields. In biology, one or more postdoctoral fellowships are quite common (indeed, routine) before an individual strikes out on his or her own. By contrast, the most typical career path for a newly graduated Ph.D. in computer science calls for appointment to a junior faculty position or a position in industry—postdoctoral fellows in computer science are relatively rare (though not unheard of).

Two foundation-supported postdoctoral programs have been influential in stimulating interest at the BioComp interface: the Sloan-Department of Energy (DOE) program and the Burroughs-Welcome program.

10.2.2.5.1 The Sloan-DOE Postdoctoral Awards for Computational Molecular Biology[9] For 8 years, the Alfred P. Sloan Foundation and the U.S. Department of Energy (Office of Biological and Environmental Research) have jointly sponsored postdoctoral research awards for scientists interested in computational molecular biology. The purpose of these fellowships has been to catalyze career transitions into computational molecular biology by those holding doctorates in mathematics, physics, computer science, chemistry, engineering or other relevant fields who would like to bring their computational sophistication to bear on the complex problems that increasingly face molecular biology.

Operationally, the program was designed to offer computationally sophisticated young scientists an intensive postdoctoral opportunity in an appropriate molecular biology laboratory. In most cases, awardees had strong educational backgrounds in a computationally intensive field, although in a few

[9]See http://www.sloan.org/programs/scitech_postdoct.shtml.

instances, awardees had backgrounds from more traditional biological orientations without the computational dimension. Of particular interest to the Sloan-DOE program are important problems in structural biology and genome analysis, including analysis of protein and nucleic acid sequence, protein and nucleic acid structure, genome structure and maps, cross-species genome analysis, multigenic traits, and structure-function relationships where the structures are from genomes, genes, or gene products.

The Sloan-DOE postdoctoral award supports up to 2 years of research in an appropriate molecular biology department or laboratory in the United States or Canada selected by the awardee. In magnitude, the award provides for a total budget of $120,000 (including indirect and overhead costs), spread over a grant period of 2 years.

10.2.2.5.2 The Burroughs-Wellcome Career Awards at the Scientific Interface[10] The Burroughs-Wellcome Career Awards at the Scientific Interface are intended to foster the early career development of researchers with backgrounds in the physical and computational sciences whose work addresses biological questions and who are dedicated to pursuing a career in academic research.[11] Prospective awardees are expected to have Ph.D.-level training in a scientific field other than biology and are encouraged to describe potential collaborations with well-established investigators working on interface problems of interest.

The program provides $500,000 over 5 years to support up to 2 years of advanced postdoctoral training and the first 3 years of a faculty appointment. In general, an awardee is expected to accept a faculty position at an institution other than the one supporting the postdoc, a requirement that is likely to spread the philosophy of interface research embodied in the program more effectively than the publishing of papers or program descriptions.

In addition, the Burroughs-Welcome Fund (BWF) requires the faculty-hiring institution to make a significant commitment to the award recipient's career development, where "significant commitment" is demonstrated by the financial and professional situation offered. Tenure-track faculty appointments are strongly preferred, accompanied by salary support and/or support for starting up a laboratory. Awardees are required to devote at least 80 percent of their time to research-related activities. Furthermore, the faculty-hiring institution must offer the awardee to take an adjunct appointment in a second department and name at least one tenured faculty member in a discipline complementary to the awardee's primary discipline who is willing to serve as an active collaborator.

10.2.2.5.3 Keck Center for Computational and Structural Biology: The Research Training Program The W.M. Keck Center for Computational and Structural Biology is an interdisciplinary and interinstitutional organization, including Baylor College of Medicine, the University of Houston, Rice University, University of Texas Health Science Center, the M.D. Anderson Cancer Center, and University of Texas Medical Branch at Galveston. Subareas of focus include computational methods and tools, biomolecular structure and function, imaging and dynamics, mathematical modeling of biosystems, and medical and genomic informatics. The faculty include some 130 members, drawn from member institutions, and a

[10]See http://www.bwfund.org/programs/interfaces/career_awards_background.html.

[11]A previous Burroughs-Wellcome Fund (BWF) program, known as Institutional Awards at the Scientific Interface, has been discontinued. (Together with the Career Awards program, it constituted the BWF Interfaces in Science effort.) The purpose of the Institutional Awards program was to support U.S. and Canadian academic institutions in developing interdisciplinary graduate and postdoctoral training programs for individuals with backgrounds in the physical, computational, or mathematical sciences to pursue biological questions. For example, pre- and postdoctoral fellows at the La Jolla Consortium and the University of Chicago's Institute for Biophysical Dynamics had to propose research projects that required the participation of two mentors—one from the quantitative sciences and one from the biological sciences—before being awarded financial support. For more on the Institutional Awards program, see N.S. Sung, J.I. Gordon, G.D. Rose, E.D. Getzoff, S.J. Kron, D. Mumford, J.N. Onuchic, et al., "Science Education: Educating Future Scientists," *Science* 301(5639):1485, 2003, available at http://www.bwfund.org/programs/interfaces/institutional_main.html.

few dozen predoctoral and postdoctoral fellows. The Keck Center was established in 1990 by a $5 million grant from the Keck foundation and currently receives more than $20 million annually in grants, from agencies such as the National Institutes of Health, the National Science Foundation, the Department of Defense, and private sources.

The Keck Center's training program, supported by the W.M. Keck Foundation and the National Library of Medicine, seeks to cross-train new scientists in both computational science and a specialized area of biology so that they can shed new light on the cellular and molecular basis of biological processes.[12] Fellowships are supported for research in algorithm development, advanced computational methods, biomedicine, crystallography, electron cryomicroscopy and computer reconstruction, genome studies, imaging and visualization, mathematical modeling of biosystems, medical informatics, neuroscience, protein dynamics and design, robotics applications in molecular biology, and the structure and function of biomolecules. The fellowship provides trainees with cross-training in computational science and in biological applications, dual mentorship, and access to cutting-edge facilities.

10.2.2.6 Faculty Retraining in Midcareer

Faculty training or retraining can augment the above opportunities. In some cases, this means participation in workshops (given release time to allow for this investment), sabbaticals spent learning a new subject, or explicitly switching from one field to another. As a rule, funded release time will be necessary to provide a break from academic constraints and to offer the time and opportunity to see biological work up close. In some cases, a good way to develop cross-disciplinary expertise is to spend a sabbatical year in the laboratory of a colleague in another discipline.

The committee was unable to find programs specifically oriented toward retraining computer scientists to do biological research. However, the National Science Foundation (NSF) does support the Interdisciplinary Grants in the Mathematical Sciences program through its Mathematical and Physical Sciences Directorate whose objective is "to enable mathematical scientists to undertake research and study in another discipline so as to expand their skills and knowledge in areas other than the mathematical sciences, subsequently apply this knowledge in their research, and enrich the educational experiences and broaden the career options of their students."[13] Recipients spend a year full-time (in a 12-month period) in a nonmathematical academic science department or in an industrial, commercial, or financial institution, and the outcome is expected to be sufficient familiarity with another discipline on the part of the supported individual "to open opportunities for effective collaboration by the mathematical scientist with researchers in another discipline." Applicants must have a tenured or tenure-track academic appointment, and the proposal must include a co-principal investigator at the level of dean (or higher-level university official) at the submitting institution as well as a commitment from the host institution or department that the hosted individual will be treated as a regular faculty member within the host unit and that at least one senior person will be provided who will serve as institutional host.

In addition, the National Institutes of Health (NIH's) National Research Service Awards program for Senior Fellows (F33) supports scientists from any field with 7 or more years of postdoctoral research experience who wish to make major changes in the direction of their research careers or who wish to broaden their scientific background by acquiring new research capabilities. In most cases, these awards are used to support sabbatical experiences for established independent scientists in which they receive training to increase their scientific capabilities. Such training must be within the scope of biomedical, behavioral, or clinical research and must offer an opportunity for individuals to broaden their scientific background or extend their potential for research in health-related areas. The maximum annual stipend is considerably lower than senior scientists typically receive, but most awardees find supplements so that they may obtain their full salaries while pursuing studies in a new field. The guidelines for eligibil-

[12]See http://cohesion.rice.edu/centersandinst/keckcenter/training.cfm?doc_id=2368.
[13]See http://www.nsf.gov/pubs/2001/nsf01115/nsf01115.htm.

ity specifically do not include previous experience in biomedical research; and thus, computer scientists would be eligible for such a program.[14]

10.2.3 Academic Organizations

Typically, academic research is conducted in departments or in centers that draw on faculty from multiple departments. The descriptions of the three departments below are simply illustrative and not exhaustive (no inference should be drawn from the fact that any given department or center is not included below):

- Cornell University maintains four distinct programs in computational biology, three hosted by a parent discipline. Biological sciences, computer science, and mathematics all offer concentrations in computational biology (the math department calls it "mathematical biology"). The only stand-alone department is the Department of Biological Statistics and Computational Biology (BSCB), a part of the College of Agriculture and Life Sciences. BSCB was originally the Department of Biometry and Biostatistics, and in 2005, it has six tenure-track faculty (plus two emeritus professors), one nontenure-track lecturer, and four "adjunct" faculty. There are 2 postdoctoral associates, 26 graduate students, and 65-70 undergraduate students. The department focuses mainly on biological statistics, computational biology, and statistical genomics. Research interests of the faculty include statistical genomics, Bayesian statistics, population genetics, epidemiology, modeling, molecular evolution, and experiment design.
- The University of California at Santa Cruz has a Department of Biomolecular Engineering, an interdisciplinary department that contains research programs in bioinformatics and experimental systems biology, among others. The bioinformatics program was originally administered by the computer engineering department. In 2005, the program has nine core tenure-track faculty members, and one affiliated faculty member. The bioinformatics curriculum includes a core of bioethics, Bayesian statistics, molecular biology, biochemistry, computational analysis of proteins, and computational genomics. Electives are drawn from biology, chemistry, computer science, and applied mathematics and statistics.
- Carnegie Mellon University (CMU) has offered programs in computational biology (through its computer science, biology, mathematics, physics, and chemistry departments) since 1989. In 2005, CMU's Department of Biological Sciences had 5 faculty involved in both computational biology and bioinformatics and genomics, proteomics, and systems biology. The department offers a B.S. in computational biology, which consists largely of a traditional biological curriculum augmented with math, programming, and computer science classes. In addition, students (B.S. or Ph.D.) can participate in the interdepartmental Merck Computational Biology and Chemistry Program, which requires students to have a home department in biology, computer science, statistics, math, or chemistry. This program was established in 1999 with a grant from the Merck Company Foundation.

Centers are often created without specific departmental affiliation because the number of departments that might plausibly contribute expertise is large. In these instances, absent a center, it is difficult to unify and coordinate research and educational activities or to convey to the outside world what the university is doing in the area. Centers are intended to be focal points for research at the BioComp interface (most often with a bioinformatics or computational biology flavor), and they usually work with departments to make new faculty appointments and provide a single point for students to learn about university programs.

Four university-based centers are described below, simply as illustrative:

[14]For more information, see http://grants.nih.gov/grants/guide/pa-files/PA-00-131.html.

• The University of California-Berkeley's Center for Integrative Genomics was founded in December 2002, supported by the Gordon and Betty Moore Foundation.[15] Its mission is to bring tools from many disciplines to bear on problems at the intersection of evolution and developmental biology. The enabling technology for new progress in this field will be acceleration of the sequencing of species genomes, and it is hoped to sequence 100 genomes of various species in the next 5 years.[16] The faculty includes 20 researchers drawn from molecular cellular biology, integrative biology, statistics, plant and microbial biology, mathematics, computer science, bioengineering, physics, paleontology, and the Lawrence Berkeley National Laboratory. The center also plans to serve an educational role, teaching or supporting the teaching of genomic science to computer science students and computer topics to biology students, as well as providing a center for graduate and postgraduate work.

• The Vanderbilt Institute for Integrative Biosystem Research and Education (VIIBRE) at Vanderbilt University (Nashville, Tennessee) was begun with an initial grant from Vanderbilt's Academic Venture Capital Fund.[17] VIIBRE has also received project-specific funding and other support from NSF, DARPA, NIH, and other institutions, enabling it to create centers of bioengineering education technologies and to begin research in cellular instrumentation and control, biomedical imaging, technology-guided therapy, biological applications of nanosystems, cellular and tissue bioengineering and biotechnology, and bioengineering education technologies. Engineers, scientists, doctors, and mathematicians conduct research for VIIBRE; more than 20 biological physics and bioengineering faculty in Vanderbilt's College of Arts and Science and the Schools of Engineering and Medicine participate in the program. VIIBRE is also developing a postdoctoral training program for physical scientists and engineers who wish to direct their careers toward the interface between biology, medicine, engineering, and the physical sciences.

• The Computational and Systems Biology Initiative (CSBi) at the Massachusetts Institute of Technology (MIT) is a campus-wide education and research program that links biologists, computer scientists, and engineers in a multidisciplinary approach to the systematic analysis of complex biological phenomena.[18] CSBi places equal emphasis on computational and experimental methods and on molecular and systems views of biological function. CSBi includes about 80 faculty members from more than 10 academic units in science, engineering, and management. Overall, membership in CSBi is self-determined, based on a self-identified interest in systems biology, and it is offered to faculty and principal investigators, postdoctoral fellows, graduate students, and research staff.

• The Institute for Biophysical Dynamics at the University of Chicago[19] is focused on interdisciplinary study of biological entities and is supported by the BWF program of Institutional Awards at the Scientific Interface. Drawing on the biological and physical science divisions of the university, the institute focuses on RNA-DNA structure, function, and regulation; protein dynamics, folding, and engineering; cytoskeleton, membranes, and organelles; hormones and cell signaling; and cell growth, death, and multicellular function. Physical scientists at the institute have expertise in macromolecular-scale manipulation via optical tweezer and chemical means; biologically relevant model systems; measurement of dynamics of macromolecules and assemblies on scales from femtoseconds to seconds; theoretical and simulation methods; soft condensed matter theory of complex and analysis of nonlinear dynamic phenomena. Part of the institute's mission is to establish cross-disciplinary training programs for students. The essential feature of the program is the placement, on a competitive basis, of predoctoral fellows with backgrounds in the physical sciences into biological science research groups, thereby

[15]See http://www.moore.org/grantees/grant_summaries_content.asp?Grantee=ucb_cig.

[16]G. Shiffrar, "New Center for Integrative Genomics to Study Major Evolutionary Changes," *College News*; see http://ls.berkeley.edu/new/02/cig.html.

[17]See http://www.vanderbilt.edu/viibre/av-goal.html and http://www.physics.vanderbilt.edu/oldpurplesite/whatshot/newsletterwinter0102.html.

[18]For more information, see http://csbi.mit.edu/whatis.

[19]For more information, see http://ibd.uchicago.edu/.

formalizing interdisciplinary connections. Fellows participate in new "translational core courses," establishing a common culture, and select an individualized program of additional coursework tailored to their research and career goals. They also take a lead role in a weekly seminar-discussion program.

Finally, in some cases centers are not associated with a specific university at all. Their purpose can be to consolidate resources on a larger scale or simply to provide a congenial intellectual home for like-minded individuals. Three nonuniversity centers are described below, again as illustrations only:

- Cold Spring Harbor Laboratory (CSHL) is a private research institution on Long Island, New York, that employs more than 800 people (300 classified as scientists) and has an annual budget of over $120 million. CSHL was established in 1889 with missions in biological research and education. In 1993, it began the annual Cold Spring Harbor Symposium on Quantitative Biology. As of 1998, it offers a Ph.D. program. Its prime research focus is on cancer biology, although it also has strong programs in plant genetics, genomics and bioinformatics, and neurobiology. In genomics, its researchers are investigating genome structure, sequencing, pattern recognition, gene expression, prediction of protein structure and function, and other related topics. A large portion of its funding comes from revenue, such as publications, intellectual property licensing, and events fees.
- The Institute for Systems Biology (ISB) is a private nonprofit institution founded in 2000 in Seattle, Washington, by Leroy Hood, Alan Aderem, and Ruedi Aebersold.[20] With a mission of applying systems biology to problems of human health such as cancer, diabetes, and diseases of the immune system, its 11 faculty members and 170 staff have expertise in fields such as immunity, proteomics, genomics, computer science, biotechnology, and biophysics. Since its founding, ISB has received its funding predominantly from federal grants, although also including private, corporate, and foundation support and industrial collaboration.[21] ISB has also spun out a number of companies to pursue commercialization opportunities around cell sorting and cancer therapies, in addition to cooperating in a multiventure capital firm-backed incubator for new biotechnology start-ups.[22] Of particular significance is the report that Hood left the University of Washington after he failed to convince it to establish a systems biology research center; he later said that he thought "the university culture and bureaucracy just could not have sufficient flexibility" to respond to the opportunity that post-Human Genome Project systems biology presented.[23]
- The Sloan-Swartz Centers for Theoretical Neurobiology were created in 1994 under the auspices of the Sloan Foundation.[24] Located at Brandeis University, California Institute of Technology, New York University, Salk Institute, and University of California, San Francisco, the Swartz Foundation also made major grants to these centers in 2000. These centers place experimentalists and theoreticians from physics, mathematics, and computer sciences in experimental brain research laboratories, where they learn about neuroscience and apply their vantage point and nontraditional skills to cooperative lines of inquiry. The centers have investigated topics such as gain fields and gain control in nerve circuits, neural coding and information theory, neural population coding and response, natural field analysis, and short-term memory.

[20]See http://www.systemsbiology.org.

[21]L. Timmerman, "Progress, Not Profit: Nonprofit Biotech Research Groups Grow in Size, Influence," *Seattle Times*, August 4, 2003.

[22]J. Cook, "Accelerator Aims to Lure, Nurture Best Ideas in Biotech," *Seattle Post-Intelligencer*, May 23, 2003.

[23]"Under Biology's Hood," *Technology Review*, September 2001, available at http://www.techreview.com/articles/01/09/qa0901.asp.

[24]See http://www.swartzneuro.org/research_a.asp and http://www.sloan.org/programs/scitech_supresearch.shtml.

10.2.4 Industry

Industrial interest in the BioComp interface is driven by the prospect of potentially very large markets in the life sciences—especially medicine. Information-enabled bioscience is further expected to create large markets for information technologies customized and adapted to the needs of life scientists—accounting in substantial measure for the interest of some large IT companies in this area. Indeed, according to the International Data Corporation, life science organizations will spend an estimated $30 billion on technology-related purchases in 2006, up from $12 billion in 2001.[25]

Life science companies (e.g., pharmaceuticals) view information technology as a (or perhaps the) key enabler for drug design and treatments that can in principle be customized to groups as small as a single individual. Consider, for example, the specific problem of finding useful organic compounds, such as drugs, to treat or reduce the effects of disease. One approach is based on the use of combinatorial methods in chemistry, genetic engineering, and high-throughput screening technology. Such an approach relies on trial-and-error to sift candidate compounds on a large scale to sidestep the complexities of data in a search for compounds with sufficient potential to be worth the effort of laboratory testing for useful outcomes; similar techniques can be used for strain improvement and natural product synthesis.[26]

A second approach is to use computational modeling and simulation. Data mining (Section 4.4.8) can be used in addition to empirical screening to identify compounds that are likely to have a desired pharmacological effect. Moreover, what the combinatorial and high-throughput empirical approach gains in expediency, it may lose in insight. For example, causality in combinatorial approaches is often difficult to attribute; and thus, it is difficult to generalize these results to other systems. Combinatorial methods are less likely to find solutions when the desired functionality is complex (e.g., when the biosynthetic route to a product is complicated or when a disease treatment relies on the inhibition, without side effects, of various pathways). Also, of course, from the standpoint of basic science, predictive understanding is at a premium. Computational simulation is thus used as the screening tool for promising compounds—a cell's predicted functional response to a given compound is used as that compound's measure of promise for further (empirical) testing. Thus, although granting drug approvals on the basis of simulations makes little sense, simulations may be able to predict with an adequate degree of reliability what drugs should not advance to expensive in vivo clinical trials.[27] Many believe that information-enabled bioscience and biotechnology have the potential to be as revolutionary as information technology was a few decades ago.

---

[25]E. Frauenheim, "Computers Replace Petri Dishes in Biological Labs," *CNET News.com,* June 2, 2003, available at http://news.com.com/2030-6679_3-998622.html?tag=fd_lede2_hed.

[26]See, for example, C. Khosla, and R.J. Zawada, "Generation of Polyketide Libraries via Combinatorial Biosynthesis," *Trends in Biotechnology* 14(9):335-341, 1996; C.R. Hutchinson, "Combinatorial Biosynthesis for New Drug Discovery," *Current Opinion in Microbiology* 1(3):319-329, 1998; A.T. Bull, A.C. Ward, and M. Goodfellow, "Search and Discovery Strategies for Biotechnology: The Paradigm Shift," *Microbiology in Molecular Biology Review* 64(3):573-606, 2000; Y. Xue and D.H. Sherman, "Biosynthesis and Combinatorial Biosynthesis of Pikromycin-related Macrolides in *Streptomyces venezuelae*," *Metabolic Engineering* 3(1):15-26, 2001; and L. Rohlin, M. Oh, and J.C. Liao, "Microbial Pathway Engineering for Industrial Processes: Evolution, Combinatorial Biosynthesis and Rational Design," *Current Opinion in Microbiology* 4(3):330-335, 2001.

[27]For example, the Tufts Center for the Study of Drug Development estimates the cost of a new prescription drug at $897 million, a figure that includes expenses of project failures (e.g., as those drugs tested that fail to prove successful in clinical trials). Since clinical trials—occurring later in the drug pipeline—are the most expensive parts of drug development, the ability to screen out drug candidates that are likely to fail in clinical trials would have enormous financial impact and would also reduce the many years associated with clinical trials. See Tufts Center for the Study of Drug Development news release, "Total Cost to Develop a New Prescription Drug, Including Cost of Post-Approval Research, Is $897 Million," May 13, 2003, available at http://csdd.tufts.edu/NewsEvents/RecentNews.asp?newsid=29. Of particular interest is a finding reported by DiMasi that if preclinical screening could increase success rates from the current 21.5 percent to 33 percent, the cost per approved drug could be reduced by $230 million (J.A. DiMasi, "The Value of Improving the Productivity of the Drug Development Process: Faster Times and Better Decisions," *PharmacoEconomics* 20(S3):1-10, 2002).

As in the case of academic organizations, specific company names provided below are illustrative and hardly exhaustive, and no inference should be drawn from the fact that any given company is not included.

10.2.4.1 Major IT Corporations

As a fast-growing, (comparatively) well-funded, and high-profile sector, life sciences research and business represents an irresistible target to large IT vendors. As such, companies such as HP and IBM have both developed suites of products and services customized for the consumption of research labs as well as the biotechnology and pharmaceutical sectors. These services are not necessarily substantially different from those that vendors provide to other sectors—a disk drive is a disk drive—but are bundled with useful software or interfaces designed with the life sciences in mind.

IBM established its Life Sciences Business Unit in 1998, incorporating hardware, consulting services, and an aggressive alliance program that includes many major vendors of bioinformatics and related software. In addition, it provides DiscoveryLink, a customized front end to IBM's successful DB/2 relational database product. Among other features, DiscoveryLink allows single-application views and queries into multiple back-end databases, providing a convenient answer to a very common situation in bioinformatics, which often deals with many databases simultaneously.

Of higher profile are IBM's research activities in computational biology. One of these is Blue Gene, the architectural successor to Deep Blue, the IBM-designed supercomputer that beat chess champion Gary Kasparov in 1997. Blue Gene, announced in 1999 as a $100 million, 5-year project, is projected to be 1 petaflop (10^{15} floating point operations per second), a thousand times more powerful than Deep Blue, and 30 times more powerful than the NEC Earth-Simulator/5120. Blue Gene is designed in part to be able to simulate the molecular forces that occur during protein folding, in order to better understand how a large protein shape emerges from a peptide sequence.[28]

Blue Gene is only one project, albeit the best known, of IBM Research's Computational Biology Center. This is a group of approximately 35 researchers who are investigating computational techniques in molecular dynamics, pattern discovery, genome annotation, heterogeneous database techniques, and so forth.

Hewlett-Packard also maintains a life sciences division, and aggressively sells hardware, software, and services to genomics research organizations, pharmaceutical companies, and agribusiness.[29] HP has had good success in winning high-profile clients.

10.2.4.2 Major Life Science Corporations

Genomic bioinformatics, and more generally the use of information technology to support research and development, has become one of the central pillars of the modern biotechnology industry, especially the pharmaceutical sector. A wave—some say a boom—of investment in bioinformatics in the late 1990s and early 2000s has tapered off, however, due to disappointing returns amid mounting costs. While few in the industry doubt the eventual impact of computational techniques, the more significant effects may not be felt for years. Even in 2002, however, corporate spending in bioinformatics was estimated to be $1 billion.[30]

The first wave of biotechnology firms, established in the 1970s, has grown into multibillion dollar operations. These firms—Amgen, Biogen, Chiron, Genentech, and Genzyme—were all founded with

[28]F. Allen, G. Almasi, W. Andreoni, D. Beece, B.J. Berne, A. Bright, J. Bruheroto, et al., "Blue Gene: A Vision for Protein Science Using a Petaflop Supercomputer," *IBM Systems Journal* 40(2):310-327, 2001, available at http://www.research.ibm.com/journal/sj/402/allen.html.

[29]See http://www.hp.com/techservers/life_sciences/overview.html.

[30]See http://www.redherring.com/investor/2002/0419/dealflop.html.

the idea of capitalizing on progress in genetic technologies. Yet because they predate the bioinformatics boom, they were often late to the game, catching up by heavy investment or by outright purchasing of other firms that had organically grown the bioinformatics capability. For example, in December of 2001, Amgen announced that it was buying the bioinformatics-rich biotech company Immunex Corp for $16 billion.[31] Genentech highlights its own bioinformatics capabilities as a key part of the research portfolio.[32] However, while these firms and the pharmaceutical giants are clearly great consumers of bioinformatics software and human resources, it is less clear to what extent they are performing original computational biology research.

A second wave of companies was founded in the 1990s, in the era of the Human Genome Project and the increase in availability of information technology. Millennium Pharmaceuticals, for example, was founded in 1993 with the goal of being a science- and technology-driven pharmaceutical company, with a capability for target discovery based on the human genome information being published. However, most of Millennium's drugs on the market have come from acquisitions, and the goal of real rational drug discovery remains challenging. Millennium does have a high-profile leader in charge of bioinformatics and uses IT for three main functions: bioinformatic inference making, such as identifying likely functions of novel proteins or the existence of gene expression patterns that correlate with disease states; chemoinformatics, searchable databases of chemical structure and biological activity; and computational analysis to predict drug candidates' physiological qualities such as absorption rates, distribution, metabolism, excretion, and toxicity.

Of higher profile is Celera, which Craig Venter founded in 1998 to compete with the publicly funded Human Genome Project. While genomics experts still argue over his methods, he certainly found innovative uses for computational and analytic techniques in stitching together the results of his "shotgun" sequencing method. Regardless of its scientific success, however, Celera has had little commercial success[33] as it turned from sequencing to the potentially more lucrative field of drug discovery. It still makes money by offering access to its proprietary databases to other biotechnology and pharmaceutical companies, but its has given up on its efforts to commercialize its software platform, selling the Celera Discovery System to sister company Applied Biosystems (both Celera and Applied Biosystems are owned by Applera Corporation). In addition to the Celera Discovery System, a subscription-based database, Applied Biosystems offers an array of software for gene sequencing, laboratory information management, and gene analysis (as well as a variety of instrumentation and reagents).

10.2.4.3 Start-up and Smaller Companies

The area still receives some attention from venture capital firms such as Flagship Ventures, Kleiner Perkins Caufield Byers, Atlas Ventures, and Alloy Ventures. However, the emphasis seems to be shifting from bioinformatics to a stronger emphasis on biology, including medical devices and drug discovery. Even companies that once positioned themselves as bioinformatics companies now describe themselves as being in the drug discovery business,[34] most notably Celera but also many smaller companies. For companies that concentrate primarily or exclusively on informatics, times are very difficult, in large part due to the same sort of bubble collapse as mainstream IT faced from 2000 onward.

Analysts blame overinvestment in the area, leading to more companies than the space can support; companies founded by IT players with insufficient biological knowledge; and increasing competition from big players such as IBM and HP.

Midsize companies such as Gene Bank and Incyte have a similar business model to Celera, offering access to proprietary databases, which often contain patented gene sequences. One model that seems to

[31]See http://www.informationweek.com/story/IWK20011221S0038.
[32]See http://www.genentech.com/gene/research/biotechnology/bioinformatics.jsp.
[33]See http://www.fool.com/portfolios/rulebreaker/2002/rulebreaker020423.htm.
[34]See http://www.bizjournals.com/washington/stories/2002/07/08/newscolumn5.html.

be more successful than others is "in silico" simulation of various biological and biomedical processes, such as offered by AnVil Informatics.[35] Beyond Genomics develops proprietary algorithms that look for large-scale biological systems such as pathways in gene and protein bioinformatics and experimental data. A second seemingly successful model is a focus on providing information about pathways and networks; Ingenuity, Cytoscape, GeneGO, PathArt, are companies that have sought to exploit this niche.

Beyond bioinformatics, vendors are attempting to develop or customize for life sciences customers a number of IT solutions, including applications for knowledge management, laboratory information management, and tracking clinical trials (including sophisticated statistical analysis).

A leading example of real computer science research being applied to biology problems is the application of distributed or grid computing to extremely computation-intensive tasks such as protein folding simulation. While many IT vendors are developing and pushing their grid platform, Stanford has been running Folding@Home, a screen-saver that anyone can download and run on a home computer, which calculates a tiny piece of the protein folding problem.[36]

10.2.5 Funding and Support

Both the federal government and private foundations support research at the BioComp interface. (The latter can be regarded as an offshoot of the historically extensive foundation support for biology research.)

10.2.5.1 General Considerations

10.2.5.1.1 The Role of Funding Institutions Funding institutions obviously exert a great deal of control and influence over the nature and direction of research. That is, researchers tend to gravitate toward research problems for which funding is available. Funding agencies can also influence the development of new talent in the field by encouraging faculty development, as illustrated non-exhaustively below:

- Release time to design new curricula and collect successful course material. However, as in peer-reviewed scientific research, the fruits of these efforts should be made public, and their successes or limitations should be openly available (e.g., as online courses or published material).
- Supervision of undergraduate special projects or research at the BioComp interface. Special projects for one or a few undergraduates (e.g., summer student projects, undergraduate theses) can be undertaken with minimal risk, and facilitating early exposure to a variety of ideas would benefit both students and faculty.
- Support for individuals who wish to make the transition to research at the BioComp interface early in their careers. Such individuals may lack the publication track record that would enable more senior researchers to undertake such a transition. Thus, support dedicated to such people may facilitate early career transitions and all of the accompanying benefits.

10.2.5.1.2 The Review Process A central dimension of funding institutions is the review process they employ to decide what research to support. Different institutions have different styles, but they all face the same types of issues.

- *Excellence*. No institution wants to support mediocre research. But as suggested below in Section 10.3.1, definitions of excellence are in many ways field-specific. Thus, an effective review process must find ways of managing this tension when proposals cross disciplinary lines.

[35] See http://www.anvilinformatics.com.

[36] See http://folding.stanford.edu. Perhaps the most famous of such distributed applications is SETI@Home, a program that supports the data processing underlying the search for extraterrestrial life.

- *Potential impact.* All else being equal, institutions would prefer to support research in which the potential impact of success is large. However, as a rule, claims of large impact are much more speculative than other claims, simply because the long-term ramifications of any given discovery are difficult to underscore in any convincing manner before the fact.
- *Technical risk.* A research investigation may or may not be successful. Research that presents the lowest technical risk (i.e., the lowest risk of failure or of being unsuccessful) is most often very closely tied to some existing and successful research. Thus, as a rule, research that is of low technical risk tends also to be of lesser potential impact.
- *Personnel risk.* Research is performed by people, and any given research effort can be executed more or less effectively depending on the people involved. Established track records of success are an important dimension of the teams proposed to undertake research but cannot be the only dimension taken into account if new researchers with good ideas are to be welcomed.
- *Budget.* Institutions with a fixed level of support to offer investigators can support a larger number of inexpensive research proposals or a smaller number of more expensive ones. All else being equal, inexpensive proposals will tend to be favored over expensive ones.

Proposals for research must weigh each of these factors and make trade-offs among them. For example, a lower budget may mean greater technical or personnel risk; a high-impact project may have greater technical risk. Funding agencies must assess the plausibility of the trade-offs that a prospective research team has made.

These notions suggest that review panels need a wide range of expertise and experience to judge the merits of new proposals effectively or to carry out peer review of scientific papers. In principle, the requisite range of expertise can be obtained through the use of a set of individual disciplinary experts whose collective expertise is adequately broad. An alternative is to use a few individuals who themselves have interdisciplinary expertise. The disadvantage of the first model is that for practical purposes it may reproduce forums in which the difficulties of cross-disciplinary understanding are manifested. The disadvantage of the second model is that such individuals may be few in number and thus difficult to enlist.

10.2.5.2 Federal Support

A variety of federal agencies support work at the BioComp interface, and this support has grown over time.

10.2.5.2.1 The National Institutes of Health For computational biology (i.e., the computing-to-biology side of the BioComp interface), the main actor in the U.S. government is the National Institutes of Health, part of the Department of Health and Human Services.

A notable instance of bioinformatics work at NIH is the National Center for Biotechnology Information (NCBI), a part of the National Library of Medicine. Established in 1988, it is NCBI that created and maintains GenBank (see Chapter 3).

The NIH's National Institute of General Medical Sciences (NIGMS) manages the Biomedical Information Science and Technology Initiative, or BISTI. BISTI represents an NIH-wide collaboration and coordination program between its many institutes and centers, as computational biology and bioinformatics activity is spread throughout the organization. In addition, NIGMS also runs the Center for Bioinformatics and Computational Biology, which focuses on theoretical and methodological infrastructure, such as modeling, simulation, theory, and analysis tools in biological networks.[37] The NIH's Center for Information Technology, in addition to providing IT services to the rest of NIH, also main-

[37]See http://www.nigms.nih.gov/news/releases/cbcb.html.

tains the Division of Computational Bioscience, which includes activities in high-performance computing and molecular modeling; it is staffed mostly by computer scientists rather than biologists and appears to focus on the computer science aspects of problems.

In addition, the National Center for Research Resources (NCRR) is a center within NIH whose mission is to create new research technologies and provide researchers access to resources such as high-end instrumentation, animal models, and cell line repositories. In FY 2004, it had a budget of slightly over a billion dollars, in large part dedicated to funding research centers, as well as individual predoctoral, postdoctoral, and career awards. NCRR's 2004-2008 strategic plan includes a number of computational biology activities within its funding programs. This includes support for software and algorithm development, mathematical modeling, and simulation. NCRR, through its Research Infrastructure Division, also supports the creation of networks to promote cross-institutional collaboration, including virtual laboratories and shared databases for a variety of specific clinical research programs. This includes the Biomedical Informatics Research Network (BIRN), an Internet2 project first funded in 2002 and slated to expand in 2004. NCRR also supports cross-discipline training at all levels of a researcher's career—for example, supporting the entry into biology of individuals with backgrounds in technical fields such as computer science, and retraining established researchers in appropriate fields.

The National Institute for Biomedical Imaging and Bioengineering (NIBIB) is the newest institute at NIH, and is unusual for its mission of assessing and developing technological capabilities for health and medical research. Its research goals and portfolio include support for a number of activities at the BioComp interface, including bioinformatics, simulation and computational modeling, image processing, brain-computer interfaces, and telemedicine. More broadly, its support for interdisciplinary training and research that draw on engineering, as well as physical and life sciences, mark it as another instrument for encouraging the development of researchers and scientists having experience with and exposure to computational science.

In addition to these institutional entities, NIH has created a set of programmatic initiatives to promote quantitative, interdisciplinary approaches to biomedical problems that involve the complex, interactive behavior of many components.[38] One initiative consists of a variety of programs to develop human capital, including those for predoctoral training for life scientists in bioinformatics and computational biology,[39] support for short courses on mathematical and statistical tools for the study of complex phenotypes and complex systems,[40] postdoctoral fellowships in quantitative biology,[41] and support for a period of supervised study and research for professionals with quantitative scientific and engineering backgrounds outside of biology or medicine who have the potential to integrate their expertise with biomedicine and develop into productive investigators.[42] The National Library of Medicine supported awards for predoctoral and postdoctoral training programs in informatics research oriented toward the life sciences (originally medical informatics but moving toward biomedical informatics in its later years).[43]

A second group of programs is targeted toward specific problems involving complex biomedical systems. This group includes an R01 program focused on genetic architecture, biological variation, and complex phenotypes (including human diseases);[44] another on quantitative approaches to the analysis of complex biological systems, with a special focus on research areas in which systems approaches are likely to result in the determination of the system-organizing principles and/or the system dynamics;[45] and still another on evolutionary mechanisms in infectious diseases.[46]

[38]See http://www.nigms.nih.gov/funding/complex_systems.html.
[39]See http://grants.nih.gov/grants/guide/pa-files/PAR-99-146.html.
[40]See http://grants.nih.gov/grants/guide/pa-files/PA-98-083.html.
[41]See http://grants.nih.gov/grants/guide/pa-files/PA-98-082.html.
[42]See http://grants.nih.gov/grants/guide/pa-files/PA-02-127.html.
[43]See http://grants.nih.gov/grants/guide/rfa-files/RFA-LM-01-001.html.
[44]See http://grants.nih.gov/grants/guide/pa-files/PA-02-110.html.
[45]See http://grants.nih.gov/grants/guide/pa-files/PA-98-077.html. This program includes P01 program project awards as well.
[46]See http://grants.nih.gov/grants/guide/pa-files/PA-02-113.html. This program includes P01 program project awards as well.

A third group of programs is institutional in nature. One program establishes new academic Centers of Excellence in Complex Biomedical Systems Research[47] that promote the analysis of the organization and dynamic behaviors of complex biological systems through the development of multi-investigator teams capable of engaging biomedical complexity with a scope of activities not possible with other funding mechanisms, including research, training, workshops, symposia, and other forms of outreach. Typical areas of interest include computationally based modeling of processes such as the cell cycle; pattern formation during embryogenesis; the flux of substrates and intermediates in metabolism; and the application of network analysis to understanding the integrated systemic host responses to trauma, burn, or other injury. A second program on Integrative and Collaborative Approaches to Research[48] encourages collaborative and integrative approaches to research on multifaceted biological problems for individual investigators with existing support who need to attract and coordinate expertise in different disciplines and approaches and require access to specialized resources, such as computational facilities, high-throughput technologies, and equipment. A third program[49] supports new quantitative approaches to the study of complex, fundamental biological processes by encouraging nontraditional collaborations across disciplinary lines through supplements to existing R01, R37, or P01 NIGMS grants to support the salary and expenses of collaborating investigators such as physicists, engineers, mathematicians, and other experts with quantitative skills relevant to the analysis of complex systems.

Finally, a major contributor to research that includes biology and computation is the NIH Roadmap. The Roadmap is a broad set of funding opportunities and programs dealing with research issues that, due to their complexity, scope, or interdisciplinary nature, could not be addressed adequately by a single NIH institute or center. Relevant BioComp programs described by the Roadmap include molecular libraries, which in part seek to develop large databases of "small molecules," and structural biology, which includes research to develop algorithmic tools for analyzing and predicting protein structure.

The most significant BioComp initiative within the Roadmap, however, is the Bioinformatics and Computational Biology program. This program seeks to create and support a National Program of Excellence in Biomedical Computing (NPEBC), a national network of software engineering and grid resources to support cutting-edge biomedical research. The prime components of the NPEBC are the National Centers for Biomedical Computing (NCBCs), seven 5-year U54 grants that total approximately $120 million, along with a larger number of R01 and R21 individual grants to support collaboration opportunities with the NCBCs.

The NCBCs are intended as more than merely well-funded research centers; their missions of training, tool creation and dissemination, community support, and liberal intellectual property policies for software and data are designed to create national networks and communities of researchers organized around BioComputational research. The structure of the grant process required the identification of three different research thrusts (or "cores"): a core of computational research, responsible for performing original work in algorithms and computer science; a core of biomedical research, or "driving biological projects," and a core Biocomputing engineering, responsible for both interfacing between computation and biomedical research, and creating the concrete tools and software systems to actualize the research.

The recipients of the first round of NCBC funding were announced in September of 2004, covering four centers. The second round, expected to fund an additional three centers, will be announced in 2005. The centers funded in the first round include:

- The Stanford Center for Physics-based Simulation of Biological Structures, an effort that seeks to create common software and algorithmic representation for modeling and simulation, addressing prob-

[47]See http://grants.nih.gov/grants/guide/rfa-files/RFA-GM-03-009.html.
[48]See http://grants.nih.gov/grants/guide/pa-files/PA-00-099.html.
[49]See http://grants.nih.gov/grants/guide/pa-files/PA-98-024.html.

lems of how to integrate models that may have widely different physical scales, have discrete or continuous approximations, or work at very different levels of abstraction. The driving biological problems for this center include RNA folding, myosin dynamics, neuromuscular dynamics, and cardiovascular mechanics.

- The National Alliance for Medical Image Computing (NAMIC) is a center based at Brigham and Women's Hospital in Boston that includes partners from universities and research centers around the country. The goal of NAMIC is to develop computational tools for analysis and visualization of image data, especially in integrating data from many different imaging technologies (e.g., magnetic resonance imaging, electroencephalography, positron emission tomography, etc.) with genomic and clinical data. The initial driving biological projects for NAMIC are various forms of neurological abnormality associated with schizophrenia.
- The Center for Computational Biology at UCLA is also investigating questions of imaging, concentrating on the production of "computational atlases," database-like structures that allow sophisticated queries of large-scale data. The computational research includes mathematics of volumes and geometry, and the driving biological projects are language development, Alzheimer's, multiple sclerosis, and schizophrenia.
- The Center for Informatics for Integrating Biology and the Bedside (I2B2), organized by a consortium of Boston-area universities, hospitals, and medical insurance providers, seeks to develop techniques to integrate and present huge sets of clinical data in ways appropriate for research into the genetic bases of disease and, thus, helping to identify appropriate targeted therapies for individual patients. This involves the development of statistical and algorithmic techniques for analyzing protein structure, as well as population dynamics. The driving biological projects include airways diseases such as asthma, hypertension, Huntington's disease, and diabetes.

10.2.5.2.2 The National Science Foundation The National Science Foundation provides a great deal of support for research at the BioComp interface through its programs of individual and institutional grants. The NSF's Directorate of Biological Sciences (BIO) formerly offered a funding program in computational biology activities. The BIO directorate ended this program in 1999,[50] not because the research no longer deserved funding, but because computational biology had "mainstreamed" to become an important part of many other biological research activities, particularly environmental biology, integrative biology, and molecular and cellular biosciences. NSF does, in its Biological Infrastructure Division, maintain a biological databases and informatics program that funds direct research into the creation of tools and datasets.

In its 2003 report *Science and Engineering Infrastructure for the 21st Century: The Role of the National Science Foundation*, NSF concludes that its support for science and engineering infrastructure (cyberinfrastructure), in which it includes next-generation computational tools and data analysis and interpretation toolkits (along with a great deal of other infrastructure elements), should increase from 22 percent of its total budget to 27 percent; it also recommends strengthening its support for cross-disciplinary fields of research. Both of these recommendations are likely to improve the funding climate for computational biology and bioinformatics, although of course they will still be competing with a number of other important infrastructure programs.

Many existing NSF funding programs emphasize interdisciplinary research and thus are effective vehicles for supporting BioComp research, although not exclusively. For example, the Integrative Graduate Education and Research Traineeship (IGERT) Program offers 5-year, $3 million grants to universities to support interdisciplinary graduate student training.[51] Many of the existing programs funded by IGERT work at the BioComp interface, such as bioinformatics, computational neuroscience, computa-

[50]See http://www.nsf.gov/pubs/1999/nsf99162/nsf99162.htm.
[51]See http://www.nsf.gov/pubs/2005/nsf05517/nsf05517.htm.

tional phylogenetics, functional genomics, and so forth. Within biology, the Frontiers in Integrative Biological Research (FIBR) program is designed to fund research projects using innovative approaches that draw on many fields, including information sciences, to attack major unanswered questions in biology.[52] It funds projects for five years at $1 million per year, and the 2005 round will fund eight projects. Also, a funding program for postdoctoral training in bioinformatics is funded at $1 million.[53]

A central and challenging application in BioComp research is an attempt to construct the entire historic phylogenetic Tree of Life. NSF is supporting this research through its Assembling the Tree of Life program, funded at $29 million; databases will contain molecular, morphological, and physiological evidence for placing taxa in relationship to other taxa. Current algorithms and data structures do not scale well at the number of taxa and data points necessary, so both computational and biological research is necessary to achieve this grand challenge.

The NSF participates with other government agencies in coordinating research agendas and programs. Of particular note is the joint initiative between the NSF Directorate for Mathematics and Physical Sciences and NIGMS to support research in mathematical biology.[54] Work supported under this initiative is expected to impact biology and advance mathematics or statistics, and the competition is designed to encourage new collaborations between the appropriate mathematical and biological scientists as well as to support existing ones. The Office of Science and Technology Policy (OSTP) included research into "molecular-level understanding of life processes" in a list of the government's top priorities for science and engineering research.[55] NSF is supporting this goal through its CAREER funding program, which is aimed at faculty members early in their careers.[56]

Finally, NSF sponsors a Small Grants Exploratory Research Program that supports high-risk research on a small scale. According to NSF, proposals eligible for support under this program must be for "small-scale, exploratory, high-risk research in the fields of science, engineering and education normally supported by NSF may be submitted to individual programs. Such research is characterized as preliminary work on untested and novel ideas; ventures into emerging research ideas; application of new expertise or new approaches to 'established' research topics; efforts having a severe urgency with regard to availability of, or access to data, facilities, or specialized equipment, including quick-response research on natural disasters and similar unanticipated events; or efforts of a similar character likely to catalyze rapid and innovative advances."[57] Typically, grants provided under this program are less than $200,000.

10.2.5.2.3 Department of Energy The Department of Energy played a key role in the initiation of the Human Genome Project. Its scientific interest was first motivated by a need to understand the biological effects of ionizing radiation, which it viewed as part of the science mission surrounding its stewardship of the nation's nuclear weapons program. Furthermore, DOE scientists have had considerable experience with advanced computation in the design and manufacturing process for nuclear weapons, a fact that DOE leveraged to investigate the genome.

Today, the Department of Energy is a major supporter of 21st century biology, because it believes that biological approaches may help it to meet its missions of energy production, global climate change mitigation, and environmental cleanup.

- For energy production, renewable energy from plants requires the design of plants with biomass that can be transformed efficiently to fuels. However, a limiting factor in developing such plants is the

[52]See http://www.nsf.gov/pubs/2004/nsf04596/nsf04596.htm.
[53]See http://www.nsf.gov/pubs/2004/nsf04539/nsf04539.html.
[54]See http://www.nsf.gov/pubs/2002/nsf02125/nsf02125.htm.
[55]See FY 2004 *Interagency Research and Development Priorities*, http://www.ostp.gov/html/ombguidmemo.pdf.
[56]See http://www.nsf.gov/pubs/2002/nsf02111/nsf02111.htm.
[57]See http://www.nsf.gov/pubs/2004/nsf042/dcletter.htm.

lack of understanding about their metabolic pathways, and knowledge of these pathways may lead to more efficient strategies for converting biomass to fuels.

- For mitigating climate change, reduction in the buildup of greenhouse gases (specifically CO_2) would be desirable. One approach to this problem is to alter natural biological cycles to store extra carbon in the terrestrial biomass, soils, and biomass that sinks to ocean depths—a sequestration approach. Research continues on the best ways to achieve large-scale carbon sequestration, and one method under investigation is tied to microbial metabolism and activities that may lead to new ways to store and monitor carbon.
- For environmental cleanup, microbes may provide a means to degrade or immobilize contaminants and accelerate the development of new, less costly strategies for cleaning up a variety of DOE waste sites. For example, microbes may be developed that can consume waste materials and degrade them or concentrate them in a form that is easier to clean up.

To address these missions, DOE supports a number of programs. Perhaps the best known is the Genomes-to-Life (GTL) program, a large research grant-providing program with four major scientific goals: (1) identification of systems of interacting proteins at the microbial level ("protein machines"), (2) characterization of gene regulatory networks, (3) exploration of microbial communities and ecosystems, and (4) development of the computational capability for modeling biological systems. To pursue these goals, the GTL program combines large experimental datasets with advanced data management, analysis, and computational simulations to create predictive simulation models of microbial function and of the protein machines and pathways that embody those behaviors. The program identifies specific challenges for computer science:[58] automated gene annotation; software to support protein expression-proteomics analysis; the ability to meaningfully and automatically extract meaning from biological technical papers; simulation for cellular networks; and model and system interoperability. These will require advances in data representation, analysis tools, integration methods, visualization techniques, models, standards, and databases. The program has funded five major projects (three at DOE labs and two at academic institutions) for a total of $103 million over the period from 2002 to 2007. In the project descriptions of the winners, four included "computational models" as part of their charge.[59]

A second DOE effort is the Microbial Genome program, which spun off from the Human Genome Project in 1994. The Microbial Genome program exploits modern sequencing technologies to sequence completely the genomes of microbes, primarily prokaryotes, based on their relevance for energy, the global carbon cycle, and bioremediation. As of April 2003, the genomes of about 100 microbes had been sequenced, most of them by the Joint Genome Institute,[60] and placed in public databases. Microbial genomics presents some particularly interesting science in that for newly sequenced microbial genomes, a large fraction of the genes identified (about 40 percent) have unknown functions and biological value. In addition, most of what is known about microbes involves microbes that are easy to culture and study or that cause serious human and animal diseases. These constitute only a small minority of all microbes living in natural environments. Most microbes are part of communities that are very difficult to study but play critical roles in Earth's ecology, and a genomic approach to understanding these microbes may be one of the only paths toward developing an understanding of them.

A third component of DOE's efforts is in structural biology. The purpose of this program is to understand the function of proteins and protein complexes that are key to the recognition and repair of DNA damage and the bioremediation of environmental contamination by metals and radionuclides. Research supported in this program focuses on determining the high-resolution three-dimensional

[58]See http://www.doegenomestolife.org/pubs/ComputerScience10exec_summ.pdf.

[59]See http://doegenomestolife.org/research/2002awards.htm.

[60]The Joint Genome Institute, established in 1997, is a consortium of scientists, engineers, and support staff from DOE's Lawrence Berkeley, Lawrence Livermore, and Los Alamos National Laboratories. See http://www.jgi.doe.gov/whoweare/index.html.

structures of key proteins; understanding the changes in protein structure related to interaction with molecules such as DNA, metals, and organic ligands; visualization of multiprotein complexes that are essential to understand DNA repair and bioremediation; prediction of protein structure and function from sequence information, and modeling of the molecular complexes formed by protein-protein or protein-nucleic acid interactions.

10.2.5.2.4 Defense Advanced Research Projects Agency With a reputation for engaging in "high-risk, high-return" research, DARPA has been a key player in the development of applications that utilize biomolecules as information processing, sensing, or structural components in anticipation of reaching the limits of Moore's law. This research area, largely supported under DARPA's biocomputation program,[61] was described in Section 8.4. Managed out of DARPA's Information Processing Technology Office (IPTO), the biocomputation program has also supported the BioSPICE program, a computational framework with analytical and modeling tools that can be used to predict and control cellular processes (described in Chapter 5 (Box 5.7)). Finally, the biocomputation program has supported work in synthetic biology (i.e., the design and fabrication of biological components and systems that do not already exist in the natural world) as well as the redesign and fabrication of existing biological systems (described in Section 8.4.2.2).

IPTO also supports a number of programs that seek to develop information technology that embodies certain biological characteristics.[62] These programs have included the following:

- *Software for distributed robotics*, to develop and demonstrate techniques to safely control, coordinate, and manage large systems of autonomous software agents. A key problem is to determine effective strategies for achieving the benefits of agent-based systems, while ensuring that self-organizing agent systems will maintain acceptable performance and security protections.
- *Mobile autonomous robot software*, to develop the software technologies needed for controlling the autonomous operation of singly autonomous, mobile robots in partially known, changing, and unpredictable environments. In this program, ideas from robot learning and control are extended, including soft computing, robot shaping, and imitation.
- *Taskable agent software kit*, to codify agent design methodology as a suite of control and decision mechanisms, to devise metrics that characterize the conditions and domain features that indicate appropriate design solutions, and to explain and formalize the notion of emergent behavior.
- *Self-regenerative systems*, to develop core technologies necessary for making computational systems able to continue operation in the face of attacks, damage, or errors. Specific avenues of investigation include biological metaphors of diversity, such as mechanisms to automatically generate a large number of different implementations of a given function that most of them will not share a given flaw; immune systems; and human cognitive models.
- *Biologically inspired cognitive architectures*, to codify a set of theories, design principles, and architectures of human cognition that are specifically grounded in psychology and neurobiology. Although implementation of such models on computers is beyond the scope of the current project, it is a natural extension once sufficiently complete models can be created.

DARPA's Defense Sciences Office (DSO) supports a variety of programs that connect biology to computing in the broad sense in which this report uses the term. These programs have included the following:

[61] See http://www.darpa.mil/ipto/programs/biocomp/index.htm.
[62] See http://www.darpa.mil/ipto/Programs/programs.htm.

- *Bio:Info:Micro.* In collaboration with DSO, IPTO and the Microsystems Technology Office, the Bio:Info:Micro program supports research in neuroprocessing and biological regulatory networks. These research thrusts seek to develop devices for interrogating and manipulating living brains and brain slices (in the neuroprocessing track) and single cells or components thereof (in the regulatory network track), and the computational tools needed to analyze and interpret information derived from these devices. Thus, neural decoding algorithms for neural spikes and local field potentials, and methods for representing spatial components in distributed systems and using decision theoretic approaches for decoding brain signals are of interest to the neuroprocessor track, and algorithms that can automatically detect patterns and networks given appropriate data and models for networks that govern cell growth and death are of interest to the regulatory track.
- *Biological input/output systems.* Focused on the design and assembly of molecular components and pathways that can be used to sense and report the presence of chemical or biological analytes, this program seeks to develop technologies to enable the facile engineering and assembly of functional biological circuits and pathways in living organisms, thereby enabling such organisms to serve as remote sentinels for those analytes. The essential notion is that the binding of an analyte to an engineered cytoplasmic or cell surface receptor will lead to regulated and specific changes in an organism, which might then be observed by imaging, spectroscopy, or DNA analysis.
- *Simulation of biomolecular microsystems.* Biological or chemical microsystems in which biomolecular sensors are integrated with electronic processing elements offer the potential for significant improvements in the speed, sensitivity, specificity, efficiency, and affordability of such systems. This program seeks to develop data, models, and algorithms for the analysis of molecular recognition processes; transduction of molecular recognition signals into measurable optical, electrical, and mechanical signals; and on-chip fluidic-molecular transport phenomena. The ultimate goal is to produce advanced computer-aided design (CAD) tools for routine analysis and design of integrated biomolecular microsystems.
- *Engineered biomolecular nanodevices and systems.* This program is focused on hybrid (biotic-abiotic) nanoscale interface technologies that enable direct, real-time conversion of biomolecular signals into electrical signals. Success in this area would enable engineered systems to exploit the high sensory sensitivity, selectivity, and efficiency that characterize many biological processes. The objective of this research is to develop hybrid biomolecular devices and systems that use biological units (e.g., protein ion channels or nanopores, g-protein-coupled receptors) for performing a sensing function but use silicon circuitry to accomplish the signal processing. Ultimately, this research is intended to lay the foundation for advanced "biology-to-digital" converter systems that enable direct, real-time conversion of biological signals into digital information.
- *Biologically inspired multifunctional dynamic robots.* This program seeks to exploit biological approaches to propulsion mechanisms for multifunctional, dynamic, energy-efficient, and autonomous robotic locomotion (e.g., running over multiple terrains, climbing trees, jumping and leaping, grasping and digging); recognition and navigation mechanisms that enable biological organisms to perform terrain following, grazing incidence landings, target location and tracking, plume tracing, and hive and swarm behavior; and the integration of these capabilities into demonstration robotic platforms.
- *Compact hybrid actuators program.* This program seeks to develop electromechanical and chemomechanical actuators that perform the same functions for engineered systems that muscle performs for animals. The performance goal is that these new actuators must exceed the specific power and power density of traditional electromagnetic- and hydraulic-based actuation systems by a factor of 10.
- *Active biological warfare sensors.* This program seeks to develop technology to place living cells with similar behavior to human cells onto chips, so that their health and behavior can be monitored for the presence of harmful chemical or biological agents.
- *Protein design processes.* This program is using two specific challenge problems to motivate research into technologies for designing novel proteins for specific biological purposes. Such design will require advances in computational models, as well as knowledge of molecular biology. The challenge

problems include tasks of designing specific proteins in less than a day that can catalyze specific chemicals or inactivate an influenza virus.

As a vehicle for pursuing its mission, DARPA typically uses Broad Agency Announcements. Some are focused on achieving a specific technical capability (e.g., a program that will develop technology for synthesizing, within 24 hours, an arbitrary 10,000-oligonucleotide sequence in quantity), whereas others are more broadly cast. In many cases, these programs target private industry as well as the more engineering-oriented academic institutions.

10.3 BARRIERS

Because work at the BioComp interface draws on different disciplines, there are barriers to effective cooperation between practitioners from each field. (In some cases, "each field" is more properly cast as the contrast between practitioners of systems biology and practitioners of empirical or experimental biology.) This section describes some of these barriers.[63]

10.3.1 Differences in Intellectual Style

It is almost axiomatic that substantial progress in any area of intellectual inquiry depends on the excellence of work undertaken in that area. On the other hand, differences in intellectual style will affect what is regarded as excellence, and computer scientists and biologists often have very different intellectual styles.

The existence of shared intellectual styles tends to increase the mutual understanding of colleagues working within their home disciplines, a fact that leads to more efficient communication and to shared epistemological understanding and commitments. However, when working across disciplines, lack of a shared intellectual style increases the difficulties for both parties in making meaningful progress.

What are some of the differences involved? While it is risky (indeed, foolhardy) to assert hard and fast differences between the disciplines, an examination of the intellectual traditions and histories associated with each discipline suggests that practitioners in each are generally socialized and educated with different styles.[64] Over time, these differences may moderate as biology becomes a more quantitative discipline (indeed, a premise of this report is that such evolution is to be encouraged and facilitated).

10.3.1.1 Historical Origins and Intellectual Traditions

Many differences in intellectual style between the two fields originate in their histories.[65] Computer science results from a marriage between mathematics and electrical engineering—although it has evolved far from these beginnings. The mathematical thread of computer science is based on formal problem statements, formulating hypotheses (conjectures) based on those statements, and generating formally correct proofs of those hypotheses. Most importantly, a single counterexample to a conjecture invalidates the conjecture. Note also that formal proofs often entail problems that are far from reality, because many real problems are simply too complex to be represented as formal problem statements that are at all comprehensible. Research in mathematics (specifically, applied mathematics) often con-

[63]An early perspective on some of these barriers can be found in K.A. Frenkel, "The Human Genome Project and Informatics: A Monumental Scientific Adventure," *Communications of the ACM* 34:40-51, 1991.

[64]An interesting ethnographic account of life in an academic biology laboratory is provided in J. Owen-Smith, "Managing Laboratory Work Through Skepticism: Processes of Evaluation and Control," *American Sociological Review* 66(3):427-452, 2001.

[65]Some of this discussion is inspired by G. Wiederhold, "Science in Two Domains," Stanford University, March 2002, updated February 2003. Unpublished manuscript.

sists of finding solutions to abstractly formulated problems and then finding real-world problems to which these solutions are applicable.

The engineering thread of computer science is based on finding useful and realizable solutions to real-world problems. The space of possible solutions is usually vast and involves different architectures and approaches to solving a given problem. Problems are generally simplified so that only the most important aspects are addressed. Economic, human, and organizational factors are at least as important as technological ones, and trade-offs among alternatives to decide on the "best" approach to solve a (simplified) problem often involve art as much as science.

Biology—the study of living things—has an intellectual tradition grounded in observation and experiment. Because biological insight has often been found in apparently insignificant information, biologists have come to place great value on data collection and analysis. In contrast to the theoretical computer scientist's idea of formal proof, biologists and other life scientists rely on empirical work to test hypotheses.

Because accommodating a large number of independent variables in an experiment is expensive, a common experimental approach (e.g., in medicine and pharmaceuticals) is to rely on randomized observations to eliminate or reduce the effect of variables that have not explicitly been represented in the model underlying the experiment. Subsequent experimental work then seeks to replicate the results of such experiments.

A biological hypothesis is regarded as "proven" or "validated" when multiple experiments indicate that the result is highly unlikely to be due to random factors. In this context, the term "proven" is somewhat misleading, as there is always some chance that the effect found is a random event. A hypothesis "validated" by experimental or empirical work is one that is regarded as sufficiently reliable as a foundation for most types of subsequent work. Generalization occurs when researchers seek to extend the study to other conditions, or when investigation is undertaken in a new environment or with more realism. Under these circumstances, the researcher is investigating whether the original hypothesis (or some modification thereof) is more broadly applicable.

Within the biological community (indeed, for researchers in any science that relies on experiment), repetition of an experiment is usually the only way to validate or generalize a finding, and replication plays a central role in the conduct of biological science. By contrast, reproducing the proof of a theorem is done by mathematicians and computer scientists mostly when a prior result is suspicious. Although there is an honored tradition of seeking alternative proofs of theorems even if the original proof is not at all suspicious, replication of results is not nearly as central to mathematics as it is to biology.

Finally, biology is constrained by nature, which makes rules (even if they are not known a priori to humans), and models of biological phenomena must be consistent with the constraints that those rules imply. By contrast, computer science is a science of the artificial—more like a game in which one can make up one's own rules—and the only "hard" constraints are those imposed by mathematical logic and consistency (hence data for most computer scientists have a very different ontological role than for biologists).

10.3.1.2 Different Approaches to Education and Training

The first introduction to computer science for many individuals involves building a computer program. The first introduction to biology for many individuals is to watch an organism grow (remember growing seeds in Dixie cups in grade school?). These differences continue in different training emphases for practitioners in computer science and biology in their undergraduate and graduate work.

To characterize these different emphases in broad strokes, formal training in computer science tends to emphasize theory, abstractions, problem solving, and formalism over experimental work (indeed, computer programming—core to the field—is itself an abstraction). Moreover, as with many mathematically oriented disciplines, much of the intellectual content of computer science is integrated and, in that sense, cumulative. By contrast, data and experimental technique play a much more central

role in a biologist's education. Traditionally, mathematics (apart from statistics) is not particularly important to biology education; indeed many biologists have entered the field because they wish to pursue science that does not involve a great deal of math. Although there is a common core of knowledge among most biologists, there is an enormous amount of highly specialized knowledge that is not tightly integrated.

A second issue, often encountered in conversion programs, is the difficulty of expanding one's horizons to choose intellectual approaches or tools appropriate to the nature of the problem. Disciplinary training in any field entails exposure to the tools and approaches of that field, which may not be the best techniques for addressing problems in another field. Thus, successful researchers and practitioners at the BioComp interface must be willing to approach problems with a wide array of methodologies and problem-solving techniques. Computer scientists often may be specialists in some specific methodology, but biological research often requires the coordination of multiple approaches. Conversely, biological labs or groups that address a wide range of questions may be more hospitable to computational researchers, because they may provide more opportunities in which computational expertise is relevant.

10.3.1.3 The Role of Theory

Theory plays a very different role and has a very different status in the two fields. For computer scientists, theoretical computer science is essentially mathematics, with all of the associated rigor, certainty, and difficulties. Of particular interest in theoretical computer science is the topic of algorithmic complexity. The most important practical results from algorithmic complexity indicate the scaling relationships between how long it takes to solve a problem and the size of the problem when its solution is based on a specific algorithm. Thus, algorithm A might solve a problem in a time of order N^2, which means that a problem that is 3 times as large would take $3^2 = 9$ times as long to solve, whereas a faster algorithm B might solve the same problem in time of order $N \log N$ (that is, $O(N \log N)$), which means that a problem 3 times as large would take $3 \log 3 = 3.29$ times as long to solve. (A specific example is that when asked to write a program to sort a list of numbers in ascending order, one of the most common programs written by novice programmers involves an $O(N^2)$ algorithm. It takes a somewhat greater degree of algorithmic sophistication to write a program that exhibits $O(N \log N)$ behavior—which can be proven to the best that is possible.)

Such results are important to algorithm design, and all computer programs embody algorithms. Depending on the functional relationship between run time and problem size, a given program that works well on a small set of test data may—or may not—work well (i.e., run in a reasonable time) for a larger set of real data. Theoretical computer science thus imposes constraints on real programs that software developers ignore at their own peril.

Computer scientists and mathematicians derive satisfaction and pleasure from elegance of reasoning, logic, and structure. Being able to explain a phenomenon or account for a dynamical behavior with a simple model is highly valued. The reason for this penchant is clear: the simpler the model, the more likely it is that the tools of analysis can be used to dissect and understand the model fully.

This sometimes means that a tendency to oversimplify overwhelms the need for preserving realistic features, to the dissatisfaction or derision of biologists. Computer scientists, of course, may well perceive a biologist's dissatisfaction as a lack of analytical or theoretical sophistication and an unwillingness to be rigorous, and often fail to recognize the complexity inherent in biological systems. In other cases, the love of elegance leads to fixation with elegant, but irrelevant, models far beyond their value outside the field, simply because the inherent model is clean and simple. In still other cases, the lack of training of computer scientists in eliciting from users the precise nature of their problems has led computer scientists to develop good solutions to problems that are not interesting to most biologists or relevant to real biological phenomena.

By contrast, many—perhaps most—biologists today have a deep skepticism about theory and

models, at least as represented by mathematics-based theory and computational models. For example, theoretical biology has a very different status within biology and has often been a poor stepchild to mainstream biology. Results from theoretical biology are often irrelevant to specific biological systems such as a particular species, and even the simplest biological organism is so complex as to render virtually impossible a theoretical analysis based on first principles. Indeed, most biologists have a long-ingrained suspicion of theoretical models that they regard as vastly oversimplified (i.e., almost all of them) and are skeptical of any purported insights that emerge from such models. (Box 10.5 provides some examples of misleading computational and mathematical models of biological phenomena.)

Box 10.5
Some Examples of Oversimplified and/or Misleading Computational and Mathematical Models in Biology

- The Turing reaction-diffusion theory for pattern formation in developmental biology—first suggested by Turing in 1952, and largely dormant until the mid-1970s, this theory, based on an activator-inhibitor system, became a focus of partial differential equations research. Initially, attempts were made to show that diffusion and reaction of the activator-inhibitor type are responsible for the development of real structures in real embryos (stripes or spots, positions of limbs and digits, etc.) However, later work has shown that the biological solution to the pattern formation problem is inelegant and "kludgy", with many "redundant" or "inefficient" parts.[1]
- A senior computer scientist faced the issue of how one might infer the structure of a genetic regulatory network from data on the presence or absence of transcription factors. In a cell, a set of genes interact to produce a protein—and the transcription factors (themselves proteins) influence the rate at which that protein is produced. His initial model of this network was a Boolean circuit, in which the presence or absence of certain factors led to the production of the protein. A typical experimental procedure in a biology lab to probe the nature of this circuit is to observe its behavior by inhibiting the production of some transcription factor and to observe whether or not the protein is produced. The analogous action in the Boolean circuit would be cutting a wire in that circuit. However, this simple analogy failed to model the actual behavior of the biological system because, in many cases, the inhibition of one transcription factor results in another set of proteins that do the same job. Thus, the notion of simple perturbation experiments that can be viewed as analogous to just snipping a wire in a logic circuit is obvious for computer scientists—but turns out to be not particularly relevant to this particular phenomenon.
- The problem of genome sequence assembly involves piecing together a large number of short sequences (fragments) into the correct master sequence. The initial computer scientist formulation of this problem was to find the shortest sequence that would contain a given set of sequences as a consecutive piece. But this formulation of the problem was completely wrong for two reasons. First, the available information on the fragments is sometime erroneous—that is, the data might indicate that a fragment would have a certain base at a given location, but in reality it would have a different base at that location. Second, DNA molecules have a great deal of repeated structure (i.e., the same sequence is typically found multiple times). Thus, the shortest sequence is not biologically plausible because that repeated structure is ignored.
- Amino acids are represented by codons (i.e., triplets of nucleotide bases). Because there are 4 nucleotides, the number of possible codons is 4^3, or 64. But for a long time, only 20 amino acids were known that occur in nature. It turns out that by assuming that the codons overlapped each other and requiring that the coding be unambiguous, only 20 codons are possible. Because of this match, a natural assumption was that an overlapping code was operative in DNA coding. However, experimental data dispelled this notion, indicating instead that multiple codons can represent the same amino acid and further that the codons were not overlapping.

[1]See, for example, G. von Dassow, E. Meir, E.M. Munro, and G.M. Odell, "The Segment Polarity Network Is a Robust Developmental Module," *Nature* 406(6792):188-192, 2000. At the same time, the reaction-diffusion approach appears to have nontrivial utility in explaining other biological phenomena, such as certain aspects of microtubule organization (C. Papaseit, N. Pochon, and J. Tabony, "Microtubule Self-organization Is Gravity-dependent," *Proceedings of the National Academy of Sciences* 97(15):8364-8368, 2000).

A related point is that computer scientists tend to assume that universal statements have no exceptions, whereas biologists have learned that there are almost always exceptions to rules. For example, if a biologist says that all crows are black, and one asks about albino crows, the answer will be, "Oh, sure, albino crows are white, but all normal crows are black." The biologist is describing the average case—all standard crows are black—but keeps in the back of his or her mind the exceptional cases. By contrast, the computer scientist wants to know if the crow database he or she is building needs to accommodate anything other than a black crow—and thus, when a computer scientist makes a biological generalization, the biologist will often jump immediately to the exceptional case as a way of dismissing the generalization.

These comments should not be taken to imply that biologists do not use theory at all; in fact, biologists use theory and models in their everyday work. The theory of evolution is among the most powerful of all scientific theories, in the sense that it underlies the scientific understanding of all natural biological phenomena. But because the outcomes of evolutionary processes are driven by a myriad of environmental and chance influences, it is difficult to make measurable or quantitative predictions about specific biological phenomena. In this context, evolution is more of an organizing principle than a predictive formalism.

Perhaps a fairer statement is that many biologists remain to be persuaded of the value of quantitative theory and abstraction on a global basis, although they accept their value in the context of specialized hypothesis, individual probes, or inquiries on a biological process. Biological researchers are beginning to see the potential explanatory value of computational and mathematical approaches—a potential that is less apparent than might be expected because of the very success of an empirical approach to biology that has been grounded in experiment and observation for many decades.

10.3.1.4 Data and Experimentation

As mentioned above, computer scientists and biologists also view data quite differently. For the computer scientist, data usually result from measurements of some computational artifact in use (e.g., how long it takes for a program to run, how many errors a program has). Because these data are tied to artifacts that have been made by human beings, they are as ephemeral and transient as the underlying artifact, which may indeed change in the next revision or release. Because computer science is a science of the artificial, the intellectual process of the computer scientist does not begin with data, but rather with an act of artifact creation, after which measurements can be taken.[66]

Indeed, for the computer scientist, the term "experimental computer science" refers to the engineering and creation of new or improved computational artifacts—hardware or software—as the central objective of intellectual efforts.[67] Engineering has intellectual biases toward model reduction, extracting key elements, and understanding subsystems in isolation before assembling larger structures. The engineering approach also rests on the idea that basic units (e.g., transistors, silicon chips) have repeatable, predictable behavior; that "modules" with specific capability (e.g., switches, oscillators, and filters) can be made from such units; and that larger systems with arbitrary complexity are, in turn, made of such modules.

In contrast, biology today is a data-driven science—and theories and models are created to fit the data. Data, presuming they are accurate, impose "hard" constraints on the biologist in much the same way that results from theoretical computer science impose hard constraints on the computer scientist. Because of the central role that data play in biology, biologists pay a great deal of attention to experi-

[66]This is not to deny that computer scientists often work with large datasets. For example, computer scientists may work with terabytes of textual or image data. But these data are the subjects of manipulation and processing, rather than being tied directly to the performance of the hardware and software artifacts of the computer scientist.

[67]National Research Council, *Academic Careers for Experimental Computer Scientists and Engineers*, National Academy Press, Washington, DC, 1994.

mental technique and laboratory procedure and instrumentation—much more so than most computer scientists pay to the comparable areas in computer science. Thus, a computer scientist with insufficient awareness of experimental design may not be accustomed to or even aware of techniques of formal model or simulation validation.

In addition, biology has not traditionally looked to engineering for insight or inspiration. For example, proteins come in an endless variety with many variations and do not necessarily have straightforward analogues to engineering parts. Experimental biologists often focus on discovering new pieces of cellular machinery and on how defective behavior stems from broken or missing pieces (e.g., mutations). Experimental work is aimed at proving or disproving specific hypotheses, such as whether or not a particular biochemical pathway is relevant to some cellular phenomena.

The training that computer scientists receive also emphasizes general solutions that give guarantees about events in terms of their worst-case performance. Biologists are interested in specific solutions that relate to very particular (although voluminous) datasets. (A further complication is that biological data are often erroneous and/or inconsistent, especially when collected in large volume.) By recognizing and exploiting special characteristics of biologically significant datasets, special-purpose solutions can be crafted that function much more effectively than general-purpose solutions. For example, in the problem of genomic sequence assembly, it turns out that by exploiting the information available concerning the size of fragments, the number of choices for where a fragment might fit is sharply restricted.

The central role that experimental data plays in biology is responsible for the fact that, to date, computer scientists have been able to make their most important contributions in areas in which the details of some biological phenomena can be neglected to some important extent. Thus, the abstraction of DNA as merely a string of characters derived from a four-letter alphabet is a very powerful notion, and considerable headway in genomics can be made knowing little else. To be sure, there are experimental errors to take into account, and a model of the noisiness of the data must be developed, but the underlying problem is pretty clear to a computer scientist.

On the other hand, as the discussion in Section 4.4.1 makes clear, there are limits to this abstraction that arise from just such "details." Also, proteomics—in which the three-dimensional structure of a protein, rather than the linear sequence, determines its function—presents even greater challenges. To understand the geometry of a three-dimensional structure, discrete mathematics—the stock in trade of the computer scientist—is far less useful than continuous mathematics.[68] Furthermore, the properties and characteristics of the specific amino acids in a protein matter a great deal to its structure and function, whereas the various nucleotide bases are more or less equivalent from an informational standpoint. In short, proteomics involves a much more substantial body of domain knowledge than does genomics.

One illustration related by a senior computer scientist working in biology is his original dream that, with enough data,[69] it would be computationally straightforward to understand the mechanisms of gene regulation. That is, with sufficient data on regulatory pathways, cascades, gene knockouts, expression levels, and their dependencies on environmental factors, how genetic regulatory networks work would become reasonable clear. With the hindsight of several years, he now believes that this dream was hopelessly naïve in that it did not account for the myriad exceptions and apparent special cases inherent in biological data that make the biologist's intellectual life very complicated indeed.

Finally, consider that many biologists are suspicious—or at least not yet persuaded—of the value and importance of high-throughput measurement of biological systems (Section 7.2). Because many biologists were educated and worked in an era in which data were scarce, experiments in biology have historically been oriented toward hypothesis testing. High-throughput data collection drives in the opposite direction,

[68]The reason is that geometric descriptions naturally involve continuous variables such as lengths and angles, and functions of those variables.

[69]Richard Karp, University of California, Berkeley, personal communication, July 29, 2002.

and relies on the ability to sift through and recognize patterns in large volumes of data whose meaning can then be inferred. Of course, predictions that emerge from the analysis of large volumes of data must still be verified one at a time, and science is today far from the point at which such analysis would, by itself, provide reliable biological conclusions. Nevertheless, such analysis can play an important role in suggesting interesting hypotheses and thus expand the options available for biological exploration.

10.3.1.5 A Caricature of Intellectual Differences

A number of one-liners that can be used to encapsulate the differences described above, though as with all one-liners, there is considerable oversimplification. Here are four:

- The goal of computer science (CS) is to develop solutions that can be useful in solving many problems, while the goal of biology is to look for solutions to individual and specific problems.
- Computer science is driven by the development of method and technique, while biology is driven by experiment and data.
- Computer scientists are trained to search for boundary conditions and constraints, whereas biologists are trained to seek signal in the noise of their experimental data.
- Computer scientists are trained to take categorical statements literally, whereas biologists use them informally.

10.3.2 Differences in Culture

Another barrier at the BioComp interface is cultural. Each field has its own cultural style, and what seems obvious to practitioners in one field may not be obvious to those in the other. Consider, for example, differences between computer science and biology. Before PowerPoint became ubiquitous to both fields, computer scientists tended to use overhead transparencies in visiting lectures, while biologists tended to use 35 mm slides. Computer science, as a discipline, can often be pursued while working at home, whereas biological lab work requires being "in the office" to a far greater extent—a computer scientist who is away from the lab may well be seen by biologists as "not being around enough" or "not being a team player." Computer scientists are accustomed to having their own office space, while biologists (especially postdoctoral associates) work out of their labs and rarely have their own offices until they achieve an appropriate seniority.

Such differences are in some sense trivial, but they do suggest the reality of different cultures, and it is helpful to explore some other differences that are not so trivial. One of the most important differences is that of intellectual style: the discussion in Section 10.3.1 would suggest that biologists (especially those untrained in quantitative sciences) may well distrust the facile approaches and oversimplified models of computer scientists or mathematicians unfamiliar with the complexities of living things, and the computer scientist may well regard the biologist as obsessed with details and molecular parts lists rather than the qualitative or quantitative whole. This section explores some issues that lie outside the domain of intellectual style.

10.3.2.1 The Nature of the Research Enterprise

When practitioners from two fields collaborate, each brings to the table the values that characterize each field. Given the importance that biologists place on the understanding of specific biological phenomena of interest, they place the highest value on answers that are specific to those phenomena. Biologists want "the answer," and they are interested in details of a computational model only insofar as they have an effect on the answer; for the most part, they care far less about a hypothetical biological phenomenon than about explaining the data obtained from experiment. Computer scientists and mathematicians, in contrast, are interested in the parameters of a model or a solution and in ways to improve

it, characterize it, understand it better, or make it more generally applicable to other problems. Thus, the biologist will likely be interested in the results of a model run on the single dataset of interest, while the computer scientist will want to run hundreds or thousands of datasets to better analyze the behavior of the model, and mathematicians will want to explore the limits of a model's applicability.[70]

An example of this cultural difference is illustrated in the history of the Gene Ontology (GO) discussed in Chapter 4. Begun in 1998 as a collaboration between researchers responsible for three model organism databases (FlyBase [Drosophila], the Saccharomyces Genome Database, and the Mouse Genome Database), GO collaborators sought to develop structured, controlled vocabularies that describe gene products in terms of their associated biological processes, cellular components, and molecular functions in a species-independent manner. In their work, these researchers have apparently not made extensive use of the (mostly domain-independent) theoretical contributions of computer science from the last 20 years, but rather have reinvented much of that work on their own. The reason for this reinvention, offered by one knowledgeable observer, is that they were unable to find computer scientists with appropriately specialized experience who were willing to sacrifice their quest for general applicability to develop a functional, usable system.[71]

A related point is that in academia, research computer scientists have very little motivation to take a software implementation beyond the prototype stage. That is, they may have developed a powerful algorithm that is likely to be useful in many biological contexts, implemented a prototype software system based on this algorithm, and convincingly demonstrated its utility in a few cases. But because most of the intellectual credit inheres in the prototype (e.g., papers for publication and promotions), research computer scientists have little motivation to move from the prototype system, which can generally be used only by those familiar with the quirks of its operation, to a more robust system that can be used by the broader community at large. Because going from prototype to broadly usable system is generally a time-intensive process, many powerful methods are not available to the biology community.

Similar considerations apply in the biology community with respect to data curation. Intellectual credit for academic biologists inheres in the publication of primary data, rather than in any long-term follow-up to ensure that the data are useful to the broader community. (Indeed, if the data are not made useful to the broader community, the researcher originally responsible for the data gains the competitive advantage of being the only one, or one of a few, able to use them.) This suggests that cultural incentives for data curation (or the lack thereof) have to be altered if data curation is to become a more significant activity in the research community.[72]

---

[70]These differences in perspective are also found at the interface of medical informatics and bioinformatics. For example, Altman notes that "the pursuit of bioinformatics and clinical informatics together is not without some difficulties. Practitioners in clinical medicine and basic science do not instantly understand the distinction between the scientific goals of their domains and the transferability of methodologies across the two domains. *They sometimes question whether informatics investigators are really devoted to the solution of scientific problems or are simply enamored of computational methodologies of unclear significance* [emphasis added]." To reduce these tensions, Altman argues—similarly to the argument presented in this report—that "informatics investigators (and their students) be able to work collaboratively with physicians and scientists in a manner that makes it clear that the creation of excellent, well-validated methods for solving problems in these domains is the paramount goal." See R.B. Altman, "The Interactions Between Clinical Informatics and Bioinformatics: A Case Study," *Journal of the American Medical Informatics Association* 7(5):439-443, 2000.

[71]Russ B. Altman, Stanford University, personal communication, December 16, 2003.

[72]One approach that has been used to support data annotation and curation activities is the data jamboree. In November 1999, the Celera Corporation hosted an invitation-only event ("the jamboree") in which participants worked for two weeks at annotating and correcting data from the *Drosophila melanogaster* genome. By all accounts a successful event that resulted in the publication of the complete sequence as well as appropriate annotations (see M.D. Adams, S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides, S.E. Scherer, et al., "The Genome Sequence of *Drosophila melanogaster*," *Science* 287(5461):2185-2195, 2000) the event featured a very informal atmosphere that promoted social connection and interaction as well as a work environment conducive to the task. The emergence of some level of community curation on Amazon and eBay may also provide some useful hints on how to proceed. In these efforts, community assessment is allowed, but there's no overall review of the quality of the assessment. Nonetheless, users have access to a diverse collection of assessments of and can do their own meta-quality control by deciding which of the reviewers to believe. This model does scale with increasing database size, although consistent curation is hardly guaranteed. It is an open question worth some investigation as to whether community commentary (perhaps supported with an appropriate technological infrastructure) could result in meaningful data curation.

10.3.2.2 Publication Venue

Although both biologists and computer scientists in academia seek to make their work known to their respective peer communities, they do so in different ways. For many areas within computer science, refereed conferences are the most prestigious places to publish leading research.[73] In biology, by contrast, conference publications are often considered less prestigious, and biologists tend to prefer journal publications. Computer scientists often write abstracts in such a way as to entice the reader to read the full paper, whereas biologists often write abstracts in such a way that the reader need not read the full paper. In publishing work that refers to a new tool, a computer scientist may be more likely to reference a Web site where the tool can be found, while a biologist may be more likely to reference a paper describing the tool.

This difference in publication venues strongly affects attempts at collaboration. Academic biologists often do not understand refereed conferences, and computer scientists often think of journals as mere repositories of papers, rather than a means of communicating results. Given that publication is the primary output of academic research, this disagreement can be very disturbing and can have important inhibitory effects on collaboration.

10.3.2.3 Organization of Human Resources

While it is clear to all parties from the start that that the professional expertise of the biologist is needed to do good work in computational biology, a view that equates computer science with programming can lead biologists to underestimate the intellectual capabilities on the computer science side necessary for computation-intensive biology problems. Thus, many biologists who do see rewards in bridging the interdisciplinary gap (especially in academia) tend to prefer doing so in their own labs, by hiring postdoctoral fellows from physics or computer science to work on their problems, keeping these ventures "in the family," rather than by establishing partnerships with more established computer scientists.

Such an approach has advantages and disadvantages. An advantage is that postdoctoral fellows with good quantitative and computational background can be exposed to the art of biological experimentation and interpretation as a part of their postdoctoral training, the result of which can be the nurturing of young interdisciplinary scientists. One disadvantage is that by engaging individuals at the beginning of their careers, the biologist is deprived of the intellectual maturity and insight that generally accompanies more seasoned computer scientists—and such maturity and insight may be most necessary for making headway on complex problems.

The integration of computational expertise into a biological research enterprise can be undertaken in different ways. For example, in some instances, a group of computer scientists can work with a group of biologists, each bringing its own computational approach to the biological problem. In other cases, a single individual with computational expertise (e.g., a postdoc) can work in an otherwise "pure" biological group, offering expertise in math, modeling, and programming. A second dimension of integration is that the bioinformaticist can play a role as a team member who engages equally with everyone else on the research team or as a "bridge" that serves as intermediate between practitioners of various disciplines. These two roles are not mutually exclusive, although the first seems to be more common.

10.3.2.4 Devaluing the Contributions of the Other

A sine qua non of interdisciplinary work is that intellectual credit be apportioned appropriately. In some cases known to the committee, the expertise of scientists from a nonbiological discipline has

[73]Such a practice arose because computer science is a fast-moving field, with a tradition of sharing discovery by online demonstration and discussion. Conferences were originally formed as a way to talk together, in person, and with relatively fast publication of results for the requirements of academia. In this context, journal publication would have been much too slow.

been used, and yet joint authorship or due credit withheld. In one story, a respected biologist held regular discussions with an excellent mathematician colleague. The biologist assigned a theoretical project on a topic already worked out by the mathematician to an in-house physics postdoctoral fellow instead of pursuing joint work with the mathematician. The results of this in-house work fell short of what was possible or desirable but displaced other serious attempts at theoretical analysis of an interesting problem.

This example suggests a view of mathematics and computer science that is ancillary and peripheral to the "real" substance of biology. The fact that computing and mathematics have developed powerful tools for the analysis of biological data makes it easy for biologists to see the computer scientist as the data equivalent of a lab technician. However, although programming is an essential dimension of most computer scientists' backgrounds, it does not follow that the primary utility of the computer scientist is to do programming. Algorithm design, to take one example, is not programming, but because algorithms must be implemented as a computer program, it is easy to confuse the two.

In other cases known to the committee, the expertise of biological scientists has been denigrated by those in computing. For example, computer scientists sometimes view a successful biological experiment as one that "merely" produces more data and do not appreciate the fundamental creative act required to devise the appropriate experiment. This attitude suggests a view of biology in which the "real" science resides in the creation of a theory or a computational model, and data are merely what is needed to populate the model.

What accounts for such attitudes? The committee believes one contributing factor is not much different than loyalty to one's discipline. Professionals in one discipline quite naturally come to believe that the ways in which they have learned to see their discipline have inherent advantages (if they did not, they would not be part of the discipline), and challenges to the intellectual paradigms they bring to the subject may well be met with a certain skepticism.

A second point to consider is that interdisciplinary work is not necessarily symmetric. This is especially true in the mix of academic research activity vis a vis applied or technical support activity. That is, it is often possible to identify one field as being the side where research advances are occurring and the other as applying some kind of support. In some cases, Ph.D.-level research in computer science can be enriched by what is routinely taught in undergraduate classes in biology, and vice versa.

For example, individuals pursuing cutting-edge research in database design may be interested in finding data models to exercise their design. They are interested in finding domain experts to help them better understand the complexities of a certain interesting problem domain, such as biology, but these database researchers see the data and the insights coming from the biologist as helping to define the problem, but as having little to do with finding the solution. Similarly, biologists may be investigating a new topic in biology and need quantitative or logistical or algorithmic help to accomplish the research, but they feel the real intellectual contribution—to biology—comes from their insights on the biological side.

The primary exception to these scenarios is where a research group in computer science gets teamed up with a research group in biological science. In such instances, the relationship can be truly symmetrical. Both parties benefit from a symbiotic relationship. Both yield practical value to the other, while gaining theoretical value for themselves. Both operate at an equivalent level of intellectual contribution. Both gain an equivalent level of real research coming out of the activity.

10.3.2.5 Attitudinal Issues

Biology laboratories are increasingly dependent on various forms of information technology. High-throughput instrumentation generates large volumes of data very quickly. Computer-based databases are the only way to keep track of a biological literature that is growing at exponential rates. Computer programs are increasingly needed to assemble and understand biological data derived from experiments or resident in databases.

With such dependence on IT, it would not be surprising if individuals who are especially knowledgeable about information technology were necessary to keep these laboratories running at high efficiency. However, computer scientists are very wary of being put into the role of technician or programmer. Computer science researchers, facing this prospect from various research disciplines, can be sensitive about wanting respect for the fundamental research advances they bring to the table.[74]

The roles of intellectual collaborator and co-equal partner are largely incompatible with the role of technician, and it is understandable that a computer scientist would want to be treated as a coequal. At the same time, a certain amount of humility and respect is also necessary. That is, the computer scientist must refrain from jumping to conclusions, must be willing to learn the facts and contemplate biological data seriously, and must not work solely on the refined abstraction problem. It may well be necessary for the computer scientist to do some mundane things to earn the confidence of the biologist partner before being able to do more interesting things.

The biologist has a role to play in facilitating partnership as well. For example, the biologist must understand that the computer scientist (especially one at the beginning of his or her career) wants to do work of publishable quality as well—work that will earn the respect of colleagues in computer science. As suggested above, programming generally does not meet this test. A second important point is to recognize without condescension the fact that many (most?) computer scientists have very little experience or familiarity with either biological concepts or data. Still a third point is the recognition that while primary data generation and experiment remain important to the life sciences, analytical work on existing data can be every bit as valuable—bioinformatics is not simply "taking someone else's data." This last point suggests a more subtle risk in partnerships—that a person with specialized skills may be regarded as a technician or a stand-alone consultant rather than as a true collaborator.

10.3.3 Barriers in Academia

One important venue for research at the BioComp interface is academia. Universities can provide infrastructure for work in this area, but institutional difficulties often arise in academic settings for work that is not traditional or easily identified with existing departments. These differences derive from the structure and culture of departments and disciplines, and lead to scientists in different disciplines having different intellectual and professional goals and experiencing different conditions for their career success. Collaborators from different disciplines must find and maintain common ground, such as agreeing on goals for a joint project, but also respect one another's separate priorities, such as having to publish in primary journals to present at particular conferences or to obtain tenure in their respective departments according to those departmental criteria. Such cross-pressures and expectations from the home departments and disciplinary colleagues remain even if the participants develop similar goals for a project.

10.3.3.1 Academic Disciplines and Departmental Structure

Universities are structured around disciplinary departments and often have considerable difficulty in supporting and sustaining interdisciplinary work. Neither fish nor fowl, the interdisciplinary researcher is often faced with the formidable task of finding an intellectual home within the university that will take the responsibility for providing tenure, research space, start-up funding, and the like. The essential problem is that a researcher working at the interface between fields X and Y is often doing work that does not fall clearly within the purview of either Department X or Department Y. When budgets are expanding and

[74]It is useful to note that research laboratories in both biology and computer science employ technicians and programmers, and such individuals serve very useful functions in each kind of laboratory. But the role of a lab technician in a biology laboratory or programmer in a computer laboratory is quite different from the role of the senior scientist who directs the biology or computer laboratory.

resources flush, it is easy for Department X or Department Y to take a risk on an interdisciplinary scholar. But as is more often the case today, when resources are scarce, each department is much more likely to want someone who fits squarely within its traditional departmental definitions, and any appointment that goes to an interdisciplinary researcher is seen as a lost opportunity.

For example, tenure letters may be requested from traditional researchers in the field for an interdisciplinary worker; despite great success, the tenure letters may well indicate that they were unfamiliar with the candidate's work. Graduate students seeking interdisciplinary training but nominally housed in a given department may have difficulty taking that department's qualifying exam, because their training is significantly different from mainstream students.

Another dimension of this problem is that publication venues often mirror departmental structures. Thus, it may be difficult to find appropriate venues for interdisciplinary work. That is, the forms of output and forums of publication for the interdisciplinary researcher may be different than for either Department X or Department Y. For example, even within computer science itself, experimental computer scientists that focus on system building often lack a track record of published papers in refereed journals, and tenure and promotion committees (often university-wide) that focus on such records for most other disciplines in the university have a hard time evaluating the worthiness of someone whose contributions have taken the form of software that the community has used extensively or presentations at refereed conferences. Even if biologists are aware in principle of such "publication" venues, they may not be aware that such conferences are heavily refereed or are sometimes regarded as the most prestigious of publication venues. Also, prestigious journals known for publishing biology research are often reluctant to devote space to papers devoted to computational technique or methodology if it does not include specific application to an important biological problem (in which case the computational dimensions are usually given a peripheral rather than primary status).

Further, the academic tenure and promotion system is biased toward individual work (i.e., work on a scale that a single individual can publish and receive credit for). However, large software systems—common in computer science and bioinformatics—are constructed by teams. Although small subsystems can be developed by single individuals, it is the whole system that provides primary value, and university-based research that is usually driven by a single-authored Ph.D. thesis or single faculty members is not very well suited to such a challenge.[75]

Finally, in most departments, it is the senior faculty that are likely to be the most influential with regard to the allocation of resources—space, tenure, personnel and research assistant support, and so on. If these faculty are relatively uninformed or disconnected from ongoing research at the BioComp interface, the needs and intellectual perspectives of interface researchers will not be fully taken into account.

10.3.3.2 Structure of Educational Programs

Stovepiping is also reflected in the structure of educational programs. Stovepiping refers to the tendency of individual disciplines to have different points of view on what to teach and how to teach it, without regard for what goes on in other disciplines. In some cases, the methods of the future are still undeveloped, or are undergoing revolution, so that suitable texts or syllabi are not yet available. Further, like individual researchers, departments tend to be territorial, protective of their realms, and insistent on ever-growing specialized course load requirements for their own students. This discourages or precludes cross-discipline shopping. Novel training creates a need for reeducation of faculty to change the design of old curricula and modernize the teaching. These changes take time and energy, and require release time from other academic burdens, whether administrative or teaching.

[75]C. Koch, "What Can Neurobiology Teach Computer Engineers?," Division of Biology and Division of Engineering and Applied Science, California Institute of Technology, January 31, 2001, position paper to National Research Council workshop, available at http://www7.nationalacademies.org/compbio_wrkshps/Christof_Koch_Position_Paper.doc.

Related to this point is the tension between breadth and depth. Should an individual trained in X who wishes to work at the intersection of X and Y undertake to learn about Y on his or her own, or seek to collaborate with an individual trained in Y? Leading-edge research in any field requires deep knowledge. But work at the interface of two disciplines draws on both of them, and it is difficult to be deep in both fields; thus, Ph.D.-level expertise in both computer science and biology may be unrealistic to expect. As a result, collaboration is likely to be necessary in all but extraordinary cases.

Thus, what is the right balance to be struck between collaboration and multiskilling of individuals? There is no hard-and-fast answer to this question, but the answer necessarily involves some of both. Even if "collaboration" with an expert in Y is the answer, the individual trained in X must be familiar enough with Y to be able to conduct a constructive dialogue with the expert in Y, asking meaningful questions and understanding answers received. At the same time, it is unlikely that an expert in X could develop in a reasonable time expertise in Y comparable to that of a specialist in Y, so some degree of collaboration will inevitably be necessary.

This generic answer has implications for education and research. In education, it suggests that students are likely to benefit from presentations by both types of expert (in X and in Y), and the knowledge that each expert has of the other's field should help to provide an integrated framework for the joint presentations. In research, it suggests that research endeavors involving multiple principal investigators (PIs) are likely to be more successful on average than single-PI endeavors.

Stovepiping can also cause problems for graduate students who are interested in dissertation work, although for graduate students these problems may be less severe than for faculty. Some universities make it easier for graduate students to do interdisciplinary work by allowing a student's doctoral work to be supervised by a committee composed of faculty from the relevant disciplines. However, in the absence of a thesis supervisor whose primary interests overlap with the graduate student's work, it is the graduate student himself or herself who must be the intellectual integrator. Such integration requires a level of intellectual maturity and perspective that is often uncommon in graduate students.

The course of graduate-level education in computing and in biology is different in some ways. In biology, students tend to propose thesis topics earlier in their graduate careers, and then spend the remainder of their time doing the proposed research. In computer science (especially more theoretical aspects), in contrast, proposals tend to come later, after much of the work has been done. Computer science graduates do not usually obtain postdoctoral positions, more commonly moving directly to industry or to a tenure-track faculty position. Receiving a postdoctoral appointment is often seen as a sign of a weak graduate experience in computer science, making postdoctoral opportunities in biology seem less attractive.

10.3.3.3 Coordination Costs

In general, the cost of coordinating research and training increases with interdisciplinary work. When computer scientists collaborate with biologists, they also are likely to belong to different departments or universities. The lack of physical proximity makes it harder for collaborators to meet, coordinate student training, and share physical resources, and studies indicate that distance has especially strong effects on interdisciplinary research.[76]

Recognizing the importance of reducing distances between collaborators, Stanford University's Bio-X program is designed specifically to foster communication campus-wide among the various disciplines in biosciences, biomedicine, and bioengineering. The Clark Center houses meeting rooms, a shared visualization chamber, low-vibration workspace, a motion laboratory, two supercomputers, the

[76]J. Cummings and S. Kiesler, *KDI Initiative: Multidisciplinary Scientific Collaborations*, report to National Science Foundation, 2003, available at http://netvis.mit.edu/papers/NSF_KDI_report.pdf; R.E. Kraut, S.R. Fussell, S.E. Brennan, and J. Seigel, "Understanding Effects of Proximity on Collaboration: Implications for Technologies to Support Remote Collaborative Work," pp. 137-162 in *Distributed Work*, P.J. Hinds and S. Kiesler, eds., MIT Press, Cambridge, MA, 2002.

small-animal imaging facility, and the Biofilms center. Other core shared facilities available to the Stanford research community include a bioinformatics facility, a magnetic resonance facility, a microarray facility, a transgenic animal facility, a cell sciences imaging facility, a product realization lab, the Stanford Center for innovation in in vivo imaging, a tissue bank, and facilities for cognitive neuroscience, mass spectrometry, electron microscopy, and fluorescence-activated cell sorting.[77]

Interdisciplinary projects are often bigger than unidisciplinary projects, and bigger projects increase coordination costs. Coordination costs are reflected in delays in project schedules, poor monitoring of progress, and an uneven distribution of information and awareness of what others in the project are doing. Coordination costs also reduce people's willingness to tolerate logistical problems that might be more tolerable in their home contexts. Furthermore, they increase the difficulty of developing mutual regard and common ground, and they lead to more misunderstandings.[78]

Coordination costs can be addressed in part through changes in technology, management, funding, and physical resources. But they can never be reduced to zero, and learning to live with greater overhead in conducting interdisciplinary work is a sine qua non for participants.

10.3.3.4 Risks of Retraining and Conversion

Retraining or conversion efforts almost always entail reduced productivity for some period of time. This fact is often viewed with dread by individuals who have developed good reputations in their original fields, and who may worry about sacrificing a promising career in their home field while entering at a disadvantage in the new one. These concerns are especially pronounced when they involve individuals in midcareer rather than recently out of graduate school.

Such fears often underlie the failure of individuals seeking to retool themselves to commit themselves fully to their new work. That is, they seek to maintain some degree of ties to their original fields—some research, some keeping up with the literature, some publishing in familiar journals, some going to familiar conferences, and so on. These efforts drain time and energy from the retraining process, but more importantly they may inhibit the necessary mind-set of success and commitment in the new domain of work. (On the other hand, keeping a foot in their old fields could also be viewed as a rational hedge against the possibility that conversion may not be successful in leading to a new field of specialization. Moreover, maintaining the discipline of continual output is a task that requires constant practice, and one's old field is likely to be the best source of such output.)

10.3.3.5 Rapid But Uneven Changes in Biology

Biology is an enormously broad field that contains dozens of subfields. Over the past few decades, these subfields have not all advanced or prospered equally. For example, molecular and cell biology have received the lion's share of biological funding and prestige, while subfields such as animal behavior or ecology have faired much less well. Molecular and cell biology (and more recently genomics, proteomics, and neuroscience) have swept through as departments modernize, in a kind of "bandwagon" effect, leaving some of the more traditional subfields to lie fallow because promising young scholars in those subfields are unable to find permanent jobs or establish their careers due to these shifts.

Moreover, prospering subfields are highly correlated with the use of information technology. Such a close association of IT with prospering fields is likely to exacerbate lingering resentments from nonprospering subfields toward the use of information technology.

---

[77]For more information see http://biox.stanford.edu/.

[78]J. Cummings and S. Kiesler, "Collaborative Research Across Disciplinary and Institutional Boundaries," *Social Studies of Science,* in press, available at http://hciresearch.hcii.cs.cmu.edu/complexcollab/pubs/paperPDFs/cummings_collaborative.pdf.

10.3.3.6 Funding Risk

Tight funding environments often engender in researchers a tendency to behave conservatively and to avoid risk. That is, unless special care is taken to encourage them in other directions (e.g., through special programs in the desired areas), researchers seeking funding are likely to pursue avenues of intellectual inquiry that are likely to succeed. Such researchers are therefore strongly motivated to pursue work that differs only marginally from previous successful work, where paths to success can largely be seen even before the actual research is undertaken. These pressures are likely to be exacerbated for senior researchers with successful and well-respected groups and hence many mouths to feed. This point is addressed further in Section 10.3.5.3.

10.3.3.7 Local Cyberinfrastructure

Section 7.1 addressed the importance of cyberinfrastructure to the biological research enterprise taken as a whole. But individual research laboratories need to be able to count on the local counterpart of community-wide cyberinfrastructure. Institutions generally provide electricity, water, and library services as part of the infrastructure that serves individual resident laboratories. But information and information technology services are increasingly as important to biological research as these more traditional services, and thus it makes sense to consider that they might be provided as a part of the local infrastructure.

On the other hand, regarding computing and information services as part of local infrastructure has institutional implications. For example, one important issue is providing centralized support for decentralized computing. Useful scientific computing must be connected to a network, and networks must interact and must be run centrally, but nonetheless, scientific computing must be accomplished in the way scientific instruments are used, that is, very much under the control of the researcher. How can institutions develop a computing infrastructure that delivers the cost effectiveness and the robustness and the reliability of well-run centralized systems while at the same time delivering the flexibility necessary to support innovative scientific use? In many research institutions, managers of centralized computing regard researchers as cowboys uninterested in exercising any discipline for the larger good, while researchers regard the managers of centralized computing as bureaucrats who are disinterested in the practice of science. Though neither of these caricatures is correct, these divergent views of how computing should effectively be deployed in a research organization will continue to exist unless the institution takes steps to reconcile them.

10.3.4 Barriers in Commerce and Business

10.3.4.1 Importance Assigned to Short-term Payoffs

In a time frame roughly coincident with the dot-com boom, commercial interest in bioinformatics was very high—perhaps euphoric in retrospect. Large, established, biotech-pharmaceutical companies, genomics-era drug discovery companies, and tiny start-ups all believed in the potential for bioinformatics to revolutionize drug design and even health care, and these beliefs were mirrored in very high stock prices.

More recently, market valuations of biotech firms have dropped along with the rest of the technology sector, and these more recent negative trends have affected the prevailing sentiment about the value of bioinformatics for drug design, at least for the short term. Although the human genome sequencing is complete, only a handful of drugs now in the pipeline stemmed from bioinformatic analysis of the genome. Bioinformatics does not automatically lead to marketable "blockbuster" drugs, and drug companies have realized that the primary bottlenecks involve biological knowledge: not enough is known of the overall biological context of gene expression and gene pathways. In the words

of one person at a 2003 seminar, "This is work for [biological] scientists, not bioinformaticists." For this reason, further large-scale business investment in bioinformatics—and indeed for any research with a long time horizon—is difficult to justify on the basis of relatively short-term returns and thus is unlikely to occur.

These comments should not be taken to imply that bioinformatics and information technology have not been useful to the pharmaceutical industry. Indeed, bioinformatics has been integrated into the entire drug development process from gene discovery to physical drug discovery, even to computer-based support for clinical trials. Also, there is a continuing belief that bioinformatics (e.g., simulations of biological systems in silico and predictive technologies) will be important to drug discovery in the long term.

10.3.4.2 Reduced Workforces

The cultural differences between life scientists and computer scientists described in Section 10.3.2 have ramifications in industry as well. For example, a sense that bioinformatics is in essence technical work or programming in a biological environment leads easily to the conclusion that the use of formally trained computer scientists is just an expensive way of gaining a year or two on the bioinformatics learning curve. After all, if all of the scientists in the company use computers and software as a matter of course and can write SQL (Structured Query Language) queries themselves, why should the company have on its payroll a dedicated bioinformaticist to serve as an interface between scientists and software? In a time of expansion and easy money, perhaps such expenditures are reasonable, but when cash must be conserved, such a person on staff seems like an expensive luxury.

10.3.4.3 Proprietary Systems

In all environments, there is often a tension between systems built in a proprietary manner and those built in an open manner, and the bioinformatics domain is no exception. Proprietary systems are often not compatible or interoperable with each other, and yet vendors often think that they can maximize revenues through the use of such systems. This tendency is particularly vexing in bioinformatics where integration and interoperability have so much value for the research enterprise. Standards and open application programming interfaces are one approach to addressing the interoperability problem. But as is often the case, many vendors support standards only to the extent that they are already incorporated into existing product lines.

10.3.4.4 Cultural Differences Between Industry and Academia

As a general rule, private industry has done better than academia in fostering and supporting interdisciplinary work. The essential reason is that disciplinary barriers tend to be lower and teamwork is emphasized when all are focused on the common goals of making profits and developing new and useful products. By contrast, the coin of the realm in academic science is individual recognition for a principal investigator as measured by his or her publication record.

This difference appears to have consequences in a variety of areas. For example, expertise related to laboratory technique is important to many areas of life sciences research. In an industrial setting, this expertise is highly valued, because individuals with such expertise are essential to the implementation of processes that lead to marketable products. These individuals receive considerable reward and recognition in an industrial setting. Although such expertise is also necessary for success in academic research, lab technicians rarely—if ever—receive rewards that are comparable to the rewards accrued by the principal investigator.

Related to this is the matter of staffing a laboratory. In today's job environment, it is common for a newly minted Ph.D. to take several postdoctoral positions. If in those positions an individual does not

develop a sufficient publication record to warrant a faculty position, he or she is for all intents and purposes out of the academic research game—a teaching position may be available, but taking a position that primarily involves teaching is not regarded as a mark of success. However, it is exactly individuals with such experience that are in many instances the backbone of industrial laboratories and provide the continuity that is needed for a product's life cycle.

The academic drive for individual recognition also tends to inhibit collaboration. Academic research laboratories can and do work together, but it is most often the case that such arrangements have to be negotiated very carefully. The same is true for large companies that collaborate with each other, but such companies are generally much larger than a single laboratory and intracompany collaboration tends to be much easier to establish. Thus, the largest projects involving the most collaborators are found in industry rather than academia.

Even "small" matters are affected by the desire for individual recognition. For example, academic laboratories often prepare reagents according to a lab-specific protocol, rather than buying standardized kits. The kit approach has the advantage of being much less expensive and faster to put into use, but often does not provide exactly the functionality that custom preparation offers. That is, the academic laboratory has arranged its processes to require such functionality, whereas an industrial laboratory has tweaked its processes to permit the use of standardized kits.

The generalization of this point is that because academic laboratories seek to differentiate themselves from each other, the default position of such laboratories is to eschew standardization of reagents, or of database structure for that matter. Standardization does occur, but it takes a special effort to do so. This default position does not facilitate interlaboratory collaboration.

10.3.5 Issues Related to Funding Policies and Review Mechanisms

As noted in Section 10.2.5.2, a variety of federal agencies support work at the BioComp interface. But the nature and scale of this support vary by agency, in terms of the procedures for making decisions about what proposals are worthy of support.

10.3.5.1 Scope of Supported Work

For example, although the NIH does support a nontrivial amount of work at the BioComp interface, its approach to most of its research portfolio, across all of its institutes and centers, focuses on hypothesis-testing research—research that investigates well-isolated biological phenomena that can be controlled or manipulated and hypotheses that can be tested in straightforward ways with existing methods. This focus is at the center of reductionist biology and has undeniably been central to much of biology's success in the past several decades.

On the other hand, the nearly exclusive focus on hypothesis testing has some important negative consequences. For example, experiments that require breakthrough approaches are unlikely to be directly supported. Just as importantly, advancing technology that could facilitate research is almost always done as a sideline. This has had a considerable chilling effect in general on what could have been, but the impact is particularly severe for implementation of computational technologies in biological sciences. That is, in effect as a cultural aspect of modern biological research, technology development to facilitate research is not considered real research and is not considered a legitimate focus of a standard grant. Thus, even computing research that would have a major impact on the advancement of biological science is rarely done (Box 10.6 provides one example of this reluctance).

It is worth noting two ironies. First, it was the Department of Energy, rather than the NIH, that supported the Human Genome Project. Second, the development of technology to conduct polymerase chain reaction (PCR)—a technology that is fundamental to a great deal of biological research today and was worthy of a Nobel Prize in 1993—would have been ineligible for funding under traditional NIH funding policy.

Box 10.6
Agencies and High-risk, High-payoff Technology Development

An example of agency reluctance to support technology development of the high-risk, high-payoff variety is offered by Robert Mullan Cook-Deegan:[1]

> In 1981, Leroy Hood and his colleagues at Caltech applied for NIH (and NSF) funding to support their efforts to automate DNA sequencing. They were turned down. Fortunately, the Weingart Institute supported the initial work that became the foundation for what is now the dominant DNA sequencing instrument on the market. By 1984, progress was sufficient to garner NSF funds that led to a prototype instrument two years later. In 1989, the newly created National Center for Human Genome Research (NCHGR) at NIH held a peer-reviewed competition for large-scale DNA sequencing. It took roughly a year to frame and announce this effort and another year to review the proposals and make final funding decisions, which is a long time in a fast-moving field. NCHGR wound up funding a proposal to use decade-old technology and an army of graduate students but rejected proposals by J. Craig Venter and Leroy Hood to do automated sequencing. Venter went on to found the privately funded Institute for Genomic Research, which has successfully sequenced the entire genomes of three microorganisms and has conducted many other successful sequencing efforts; Hood's groups, first at Caltech and then at the University of Washington, went on to sequence the T cell receptor region, which is among the largest contiguously sequenced expanses of human DNA. Meanwhile, the army of graduate students has yet [in 1996, eds.] to complete its sequencing of the bacterium *Escherichia coli*.

[1]R. Mullan Cook-Deegan, "Does NIH Need a DARPA?," *Issues in Science and Technology* XIII:25-28, Winter 1996.

To illustrate the consequences in more concrete but future-oriented terms, the list below suggests some of the activities that would be excluded under a funding model that focuses only on hypothesis-testing research:

- Developing technologies that enable data collection from a myriad of instruments and sensors, including real-time information about biological processes and systems, that permit us to refine and annotate this information and incorporate it into accessible repositories to facilitate scientific study or biomedical procedures;
- Flexible database systems that allow incorporation of multiscale, multimodal information about biological systems by enabling the inclusion (by data federation techniques such as mediation) of information distributed in an unlimited number of other databases, data collections, Web sites and so on;
- Acquisition of "discovery-driven" data (discovery science, as described in Chapter 2) to populate datasets useful for computational analytical methods, or improvements in data acquisition technology and methodology that serve this end;
- Development of new computational approaches to meet challenges of complex biological systems (e.g., improved algorithmic efficiency, development of appropriate signal processing or signal detection statistical approaches to biological data); and
- Data curation efforts to correct and annotate already-acquired data to facilitate greater interoperability.

These considerations suggest that expanding the notion of hypothesis may be useful. That is, the discussion above regarding hypothesis testing refers to *biological* hypotheses. But to the extent that the kinds of research described in the immediately preceding list are in fact part of 21st century biology, nonbiological hypotheses may still lead to important biological discoveries. In particular, a plausible and well-supported *computational* hypothesis may be as important as a biological one and may be instrumental in advancing biological science.

Today, a biological research proposal with excellent computational hypotheses may still be rejected because reviewers fail to see a clearly articulated biological hypothesis. To guard against such situa-

tions, funding agencies and organizations would be well served by including in the review process reviewers with the expertise to identify plausible and well-supported computational hypotheses that may aid their biological colleagues in reaching a sound and unbiased conclusion about research proposals at the BioComp interface.

More generally, these considerations involve changing the value proposition for what research dollars should support. At an early point in a research field's development, it certainly makes sense to emphasize very strongly the creation of basic knowledge. But as a field develops and evolves, it is not surprising that a need to consolidate knowledge and make it more usable begins to emerge. In the future, a new balance will have to be struck between the creation of new knowledge and making that knowledge more valuable to the scientific community.

10.3.5.2 Scale of Supported Work

In times of limited resources (and times of limited resources are always with us), unconventional proposals are suspect. Unconventional proposals are even more suspect when they require large amounts of money. No better example can be found than the reactions in many parts of the life sciences research community to the Human Genome Project when it was first proposed—with a projected price tag in the billions of dollars, the fear was palpable that the project would drain away a significant fraction of the resources available for biological research.[79]

Work at the BioComp interface, especially in the direction of integrating state-of-the-art computing and information technology into biological research, may well call for support at levels above those required for more traditional biology research. For example, a research project with senior expertise in both biology and computing may well call for support for co-principal investigators. Just as biological laboratories generally require support for lab technicians, a BioComp project could reasonably call for programmers and/or system administrators. (A related point is that for a number of years in the recent past [i.e., during the dot-com boom years] computer scientists commanded relatively high salaries.)

In addition, some areas of modern life sciences research, such as molecular biology, rely on large grants for the purchase of experimental instruments. The financial needs for instrumentation and laboratory equipment to collect the data necessary for undertake the data-intensive studies of 21st century biology are significant, and are often at a scale that is unaffordable to all but a small number of academic institutions. Although large grants are not unheard of in computer science, the across-the-board dependence of important subfields of biology on experiment means that a larger fraction of biological research is supported through such mechanisms than is true in computer science.

To the extent that proposals for work at the BioComp interface are more costly than traditional proposals and supported by the same agencies that fund those traditional proposals, it will not be surprising to find resistance when they are first proposed.

What is the scale of increased cost that might be associated with greater integration of information technology into the biological research enterprise? If one believes, as does the committee, that information technology will be as transformative to biology as it has been to many modern businesses, IT will affect the way that biological research is undertaken and the discoveries that are made, the infrastructure necessary to allow the work to be done, and the social structures and organizations necessary to support the work appropriately.

Similar transformations have occurred in fields such as high finance, transportation, publishing, manufacturing, and discount retailing. Businesses in these fields tend to invest 5-10 percent of their gross revenues in information technology,[80] and this is with data that is well structured and understood. It is thus not unreasonable to suggest that a full integration of information technology into the biological research enterprise might have a comparable cost. Today, there is federal support for only a very small fraction of that amount.

---

[79]See, for example, L. Roberts, "Controversial from the Start," *Science* 291(5507):1182-1188, 2001.

[80]See, for example, http://www.bain.com/bainweb/publications/printer_ready.asp?id=17269.

10.3.5.3 The Review Process

Within the U.S. government, there are two styles of review. In the approach relying mainly on peer review (used primarily by NIH and NSF), a proposal is evaluated by a review panel that judges its merits, and the consensus of the review panel is the primary factor that influencing a decision that a proposal does or does not merit funding. When program budgets are limited, as they usually are, the program officer decides on actual awards from the pool of proposals designated as merit-worthy. In the approach relying on program officer judgment (used primarily by DARPA), a proposal is generally reviewed by a group of experts, but decisions about funding are made primarily by the program officer.

The dominant style of review mechanism in agencies that support life sciences research is peer review. Peer review is intended as a method of ensuring the soundness of the science underlying a proposal, and yet it has disadvantages. To quote an NRC report,[81]

> The current peer-review mechanism for extramural investigator-initiated projects has served biomedical science well for many decades and will continue to serve the interests of science and health in the decades to come. NIH is justifiably proud of the peer review mechanism it has put in place and improved over the years, which allows detailed independent consideration of proposal quality and provides accountability for the use of funds. However, any system that focuses on accountability and high success rates in research outcomes may also be open to criticism for discriminating against novel, high-risk proposals that are not backed up with extensive preliminary data and whose outcomes are highly uncertain. The problem is that high-risk proposals, which may have the potential to produce quantum leaps in discovery, do not fare well in a review system that is driven toward conservatism by a desire to maximize results in the face of limited funding resources, large numbers of competing investigators, and considerations of accountability and equity. In addition, conservatism inevitably places a premium on investing in scientists who are known; thus there can be a bias against young investigators.

Almost by definition, peer review panels are also not particularly well suited to considering areas of research outside their foci. That is, peer review panels include the individuals that they do precisely because those individuals are highly regarded as experts within their specialties. Thus, an interdisciplinary proposal that draws on two or more fields is likely to contain components that a review panel in a single field is not able to evaluate as well as those components that do fall into the panel's field.

A number of proposals have been advanced to support a track of scientific review outside the standard peer review panels. For example, the NRC report recommended that NIH establish a special projects program located in the office of the NIH director, funded at a level of $100 million initially to increase over a period of 10 years to $1 billion a year, whose goal would be to foster the conduct of innovative, high-risk research. Most importantly, the proposal calls for a set of program managers to select and manage the projects supported under this program. These program managers would be characterized primarily by an outstanding ability to develop or recognize unusual concepts and approaches to scientific problems. Review panels constituted outside the standard peer review mechanisms and specifically charged with the selection of high-risk, high-payoff projects would provide advice and input to program managers, but decisions would remain with the program managers. Research initially funded through the special projects program that generated useful results would be handed off after 3-5 years for further development and funding through standard NIH peer review mechanisms. Whether this proposal, or a similar one, will be adopted remains to be seen.

Different agencies also have different approaches to the proposals they seek. For example, agencies differ in the amount of detail that they insist potential grantees provide in these proposals. Depending on the nature of the grant or contract sought, one agency might require only a short proposal of a few pages and minimal documentation, whereas another agency might require many more pages, insisting on substantial preliminary results and extensive documentation. An individual familiar with one kind

[81]National Research Council, *Enhancing the Vitality of the National Institutes of Health: Organizational Change to Meet New Challenges*, The National Academies Press, Washington, DC, 2003, p. 93.

of approach may not be able to cope easily with the other, and the overhead involved in coping with an unfamiliar approach can be considerable.

As one illustration, the committee heard from a professor of computer science, accustomed to the NSF approach to proposal writing, who reported that while many biology departments have grant administrators who provide significant assistance in the preparation of proposals to NIH (e.g., telling the PI what is required, drafting budgets, filling out forms, submitting the proposal), his department (of computer science) was unable to provide any such assistance—and indeed lacked anyone at all with expertise in the NIH proposal process. As a result, he found the process of applying for NIH support much more onerous than he had expected.

10.3.6 Issues Related to Intellectual Property and Publication Credit

Issues related to intellectual property (IP) are largely outside the scope of this report. However, it is helpful to flag certain IP issues that are particularly likely to be relevant in advancing the frontiers at the intersection of computer science and biology. Specifically, because information technology enables the sensible use of enormous volumes of biological data, biological findings or results that emerge from such large volumes are likely to involve the data collection work of many parties (e.g., different labs). Indeed, biology as a field recognizes as significant, and even primary, the generation of good experimental data about biological phenomena. By contrast, multiparty collaborations on a comparable scale are unusual in the world of computer science, and datasets themselves are less significant. Thus, computer scientists may well be taken aback by the difficulties in negotiating permissions and credit.

A second issue arises that is related to tensions between open academic research and proprietary commercialization of intellectual advantages. Because of the potential that advances in bioinformatics will have great commercial value, there are incentives to keep some research in bioinformatics proprietary (hence, not easily accessible to the peer community, less amenable to peer review, and less relevant to professional development and advancement). In principle, this is not particularly different at the BioComp interface than in any other research area of commercial value. Nevertheless, the fact that traditions and practices from two different disciplines (disciplines that are at the forefront of economic growth today) are involved rather than just one may exacerbate these tensions.

A third point is the potential tension between making data publicly available and the intellectual property rights of journal publishers. For example, some years ago a part of the neuroscience community sought to build a functional positron emission tomography database. In the course of their efforts, they found that they needed to add substantial prose commentary to the image database to make it useful. Some of the relevant neuroscience journals were reluctant to give permission to use large extracts from publications in the database. To the extent that this example can be generalized, it suggests that efforts to build a far-reaching cyberinfrastructure for biology will have to identify and deal with intellectual property issues as they arise.[82]

[82]In responding to this report in draft, a reviewer argued that by taking collective action, the major research institutions could exert strong leverage on publishers to relax their copyright requirements. Today, many top-rated journals require as a condition of publication the transfer of all copyright rights from the author to the publisher. Given the status of these journals, this reviewer argued that it is a rare researcher who will take his or her paper from a top-rated journal to a secondary journal with less stringent requirements in order to retain copyright. However, the researcher's home institution could adopt a policy in which the institution retained the basic copyright (e.g., under the work-for-hire provisions of current copyright law) but allowed researchers to license their work to publishers but not to transfer the copyright on their own accord. Under such circumstances, goes the argument, journal publishers would be faced with a situation of rejecting work not just from one researcher but from all researchers at institutions with such a policy—a situation that would place far more pressure on journal publishers to relax their requirements and would improve the ability of researchers to share their information through digital resources and databases. The committee makes no judgment about the wisdom of this approach, but believes that the idea is worth mention.